



500.43155X00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s): N. Watanabe

Serial No.: 10/669,325

Filed: September 25, 2003

Title: REMOTE COPY SYSTEM

LETTER CLAIMING RIGHT OF PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

November 6, 2003

Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, the applicant(s) hereby
claim(s) the right of priority based on:

Japanese Patent Application No. 2003-205617
Filed: August 4, 2003

A certified copy of said Japanese Patent Application is attached.

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP

Carl I. Brundidge

Registration No.: 29,621

CIB/rr
Attachment

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 3 年 8 月 4 日
Date of Application:

出 願 番 号 特 願 2 0 0 3 - 2 0 5 6 1 7
Application Number:
[ST. 10/C]: [J P 2 0 0 3 - 2 0 5 6 1 7]

出 願 人 株式会社日立製作所
Applicant(s):

2 0 0 3 年 1 0 月 2 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康 夫

出証番号 出証特 2 0 0 3 - 3 0 8 1 2 0 6

【書類名】 特許願

【整理番号】 K03008941A

【あて先】 特許庁長官殿

【国際特許分類】 G06F 12/00

【発明者】

 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所システム開発研究所内

 【氏名】 渡辺 直企

【特許出願人】

 【識別番号】 000005108

 【氏名又は名称】 株式会社 日立製作所

【代理人】

 【識別番号】 100075096

 【弁理士】

 【氏名又は名称】 作田 康夫

【手数料の表示】

 【予納台帳番号】 013088

 【納付金額】 21,000円

【提出物件の目録】

 【物件名】 明細書 1

 【物件名】 図面 1

 【物件名】 要約書 1

【プルーフの要否】 要

【書類名】 明細書**【発明の名称】 リモートコピーシステム****【特許請求の範囲】****【請求項 1】**

計算機に接続され、第1のディスク装置と第1のコントローラとを有する第1の記憶装置システムと、

第2のディスク装置と第2のコントローラとを有する第2の記憶装置システムと

、

前記第1の記憶装置システム及び前記第2の記憶装置システムとに接続される、第3の記憶領域と第3のコントローラとを有する第3の記憶装置システムとを有するシステムであって、

前記第1のコントローラは、

前記計算機から受信したライト要求に応じて、前記計算機から受信したライトデータと該ライトデータが書き込まれる位置を示すアドレス情報とを含むジャーナルを前記第3の記憶装置システムに送信し、

前記ライトデータを前記第1のディスク装置に格納し、

前記ジャーナルの送信後に前記計算機に前記ライト要求に対する応答を返し、

前記第2のコントローラは、

前記第1のコントローラによって発行され、前記ジャーナルを前記第2の記憶装置システムが取得する際に使用される、前記ジャーナルの格納位置を含む第1の制御情報を受信し、

該第1の制御情報に基づいて、前記ジャーナルを前記第3の記憶サブシステムから取得し、

前記ジャーナルに含まれるアドレス情報に基づいて前記ライトデータを前記第2のディスク装置に格納することを特徴とするシステム。

【請求項 2】

請求項 1 記載のシステムにおいて、

前記第1の記憶装置システムは、前記第1の制御情報を格納しており、

前記第2のコントローラは、前記ジャーナルを取得した後、その旨を示す第2の

制御情報発行し、

前記第1のコントローラは、前記第2の制御情報を取得した後、前記第1の記憶装置システム内に格納されている前記第1の制御情報を廃棄可能とすることを特徴とするシステム。

【請求項 3】

請求項2記載のシステムにおいて、

前記第1のコントローラは前記第3の記憶装置システムに前記第1の制御情報を送信し、前記第2のコントローラは前記第3の記憶領域に格納された前記第1の制御情報を前記第3の記憶装置システムから取得し、

前記第2のコントローラは前記第2の制御情報を前記第3の記憶装置システムに送信し、前記第1のコントローラは前記第3の記憶領域に格納された前記第2の制御情報を前記第3の記憶装置システムから取得することを特徴とするシステム。

【請求項 4】

請求項 3 記載のシステムにおいて、

前記第3の記憶装置システムは、前記第1の制御情報及び前記第2の制御情報を前記第3の記憶領域内の各々異なる論理ボリュームに格納し、論理ボリュームごとに前記第1の記憶装置システムもしくは前記第2の記憶装置システムのいずれからの書き込み要求を許可するかを設定することを特徴とするシステム。

【請求項 5】

請求項 3 記載のシステムにおいて、

前記第1の記憶装置システムに障害が生じた場合に、前記第2の記憶装置システムは、前記第3の記憶装置システムに格納されている第1の制御情報を参照して、前記第2のディスク装置に格納されていないライトデータを有するジャーナルを前記第 3 の記憶装置システムから取得し、取得したジャーナルに含まれるアドレス情報に基づいて、取得したジャーナルに含まれるライトデータを前記第2のディスク装置に格納することを特徴とするシステム。

【請求項 6】

請求項 5 記載のシステムにおいて、

前記第2の記憶装置システムは、前記第1の記憶装置システムに障害が生じた後

に、前記第2の記憶装置システムに接続される計算機からライト要求を受信した場合に、該ライト要求に従って書き込まれるライトデータの格納位置を示す差分情報を有しており、

前記第1の記憶装置システムが障害から回復した場合に、前記第2の記憶装置システムは、前記差分情報が示す格納位置に格納されているデータを、前記第1の記憶装置システムと前記第2の記憶装置システムとを接続する通信路を介して、前記前記第1の記憶装置システムに送信することを特徴とするシステム。

【請求項7】

請求項2記載のシステムにおいて、

更に前記第1の記憶装置システムと前記第2の記憶装置システムとに接続される通信路を有し、

前記第1のコントローラは前記第1の制御情報を前記通信路を介して前記第2の記憶装置システムに送信し、

前記第2のコントローラは前記第2の制御情報を前記通信路を介して前記第1の記憶装置システムに送信することを特徴とするシステム。

【請求項8】

請求項7記載のシステムにおいて、

前記第3の記憶装置システムに障害が生じた場合に、前記第1のコントローラは計算機から受信するライトデータを前記通信路を介して前記第2のコントローラに送信し、前記第2のコントローラは前記通信路を介して受信したライトデータを前記第2のディスク装置に格納することを特徴とするシステム。

【請求項9】

請求項3記載のシステムにおいて、

更に前記第1の記憶装置システムと前記第2の記憶装置システムとに接続される第4の記憶装置システムを有し、

前記第3の記憶装置システムに障害が発生した場合に、前記第1の記憶装置システムと前記第2の記憶装置システムは、前記第4の記憶装置システムを介して、ジャーナル、第1の制御情報、及び第2の制御情報を送受信することを特徴とするシステム。

【請求項 10】

請求項 3 記載のシステムにおいて、

更に前記第 1 の記憶装置システムと前記第 2 の記憶装置システムとに接続される第 4 の記憶装置システムを有し、

前記第 1 のコントローラは時刻情報を有するジャーナルを前記第 3 の記憶装置システム若しくは前記第 4 の記憶装置システムのいずれかに送信し、

前記第 2 のコントローラはジャーナルを前記第 3 の記憶装置システム及び前記第 4 の記憶装置システムから取得して、取得したジャーナルに含まれるライトデータを、該ジャーナルに付与された時刻情報が示す時刻順に前記第 2 のディスク装置に書き込むことを特徴とするシステム。

【請求項 11】

計算機に接続され第 1 のディスク装置を有する第 1 の記憶装置システムと、第 2 のディスク装置を有する第 2 の記憶装置システムと、前記第 1 の記憶装置システム及び前記第 2 の記憶装置システムとに接続される第 3 の記憶装置システムとを有するシステムにおいて、前記第 1 の記憶装置システム、前記第 2 の記憶装置システム、及び前記第 3 の記憶装置システム間で実行されるリモートコピー方法であって、

前記第 1 の記憶装置システムが前記計算機からライト要求とライトデータとを受信するライト要求受信ステップと、

前記第 1 の記憶装置システムが、前記ライトデータと前記ライト要求に含まれるアドレス情報とを有するジャーナルを前記第 3 の記憶装置システムに書き込むジャーナル書き込みステップと、

前記ジャーナルを前記第 2 の記憶装置システムが読み出すために必要な、前記ジャーナルの格納位置を含む第 1 の制御情報を、前記第 1 の記憶装置システムが発行する第 1 制御情報発行ステップと、

前記第 2 の記憶装置システムが、前記第 1 の制御情報を取得する、第 1 制御情報取得ステップと、

前記第 2 の記憶装置システムが、前記第 1 の制御情報に基づいて、前記ジャーナルを読み出すジャーナル読み出しステップと、

前記第2の記憶装置システムが、前記ジャーナルに含まれるアドレス情報に従って、前記ジャーナルに含まれるライトデータを該第2の記憶装置システムが有するディスク装置に格納するライトデータ書き込みステップとを有することを特徴とするリモートコピー方法。

【請求項12】

請求項11記載のリモートコピー方法において、

前記第1の記憶装置システムは、前記第1の制御情報を格納しており、

更に、前記ジャーナル読み出しステップ後、前記第2の記憶装置システムが該ジャーナルを読み出した旨を示す第2の制御情報発行する第2制御情報発行ステップと、

前記第1の記憶装置システムが、前記第2の制御情報を受信する第2制御情報受信ステップとを有し、

前記第1の記憶装置システムは前記第2の制御情報を受信した後に前記第1の制御情報を破棄することを特徴とするリモートコピー方法。

【請求項13】

請求項11記載のリモートコピー方法において、

前記第1制御情報発行ステップは、前記第1の制御情報を前記第3の記憶装置システムに書き込むステップを有し、

前記第1制御情報取得ステップは、前記第3の記憶装置システムから前記第1の制御情報を読み出すステップを有することを特徴とするリモートコピー方法。

【請求項14】

請求項13記載のリモートコピー方法において、

前記第3の記憶装置システムは、前記第1の制御情報及び前記第2の制御情報を前記第3の記憶領域内の各々異なる論理ボリュームに格納し、論理ボリュームごとに前記第1の記憶装置システムもしくは前記第2の記憶装置システムのいずれからの書き込み要求を許可するかを設定することを特徴とするリモートコピー方法。

【請求項15】

請求項13記載のリモートコピー方法において、更に、

前記第1の記憶装置システムに障害が生じた場合に、前記第2の記憶装置システムが、前記第3の記憶装置システムに格納されている第1の制御情報を参照して、該第2の記憶装置システムが有するディスク装置に格納されていないライトデータを有するジャーナルを前記第3の記憶装置システムから取得するステップと、取得したジャーナルに含まれるアドレス情報に基づいて、取得したジャーナルに含まれるライトデータを前記第2の記憶装置システムが有するディスク装置に格納するステップとを有することを特徴とするリモートコピー方法。

【請求項 16】

請求項 15 記載のリモートコピー方法において、

前記第1の記憶装置システムに障害が生じた後、前記第2の記憶装置システムが前記第2の記憶装置システムに接続される計算機からライト要求を受信するステップと、

前記第2の記憶装置システムが、前記ライト要求に従って書き込まれるライトデータの格納位置を示す差分情報保持するステップと、

前記第1の記憶装置システムが障害から回復した場合に、前記差分情報が示す格納位置に格納されているデータを、前記第1の記憶装置システムと前記第2の記憶装置システムとを接続する通信路を介して、前記前記第1の記憶装置システムに送信するステップとを有することを特徴とするリモートコピー方法。

【請求項 17】

請求項 11 記載のリモートコピー方法において、

前記第1の記憶装置システムと前記第2の記憶装置システムとは通信路によって接続されており、

前記第1制御情報発行ステップは、前記第1の記憶装置システムが前記第1の制御情報を前記通信路を介して前記第2の記憶装置システムに送信するステップを有し、

前記第1制御情報取得ステップは、前記第2の記憶装置システムが前記第1の制御情報を前記通信路を介して前記第1の記憶装置システムから受信するステップを有することを特徴とするリモートコピー方法。

【請求項 18】

請求項 17 記載のリモートコピー方法において、前記第 3 の記憶装置システムに障害が生じた場合に更に、

前記第 1 の記憶装置システムは前記計算機から受信するライトデータを前記通信路を経由して前記第 2 の記憶装置システムに送信するステップと、

前記第 2 の記憶装置システムが前記通信路を経由して受信したライトデータを該第 2 の記憶装置システムが有するディスク装置に格納するステップとを有することを特徴とするリモートコピー方法。

【請求項 19】

請求項 13 記載のリモートコピー方法において、

前記システムは更に、前記第 1 の記憶装置システムと前記第 2 の記憶装置システムとに接続される第 4 の記憶装置システムを有し、

前記第 3 の記憶装置システムに障害が発生した場合に、前記第 1 の記憶装置システムと前記第 2 の記憶装置システムは、前記第 4 の記憶装置システムを介して、ジャーナル及び第 1 の制御情報を送受信することを特徴とするリモートコピー方法。

【請求項 20】

請求項 13 記載のリモートコピー方法において、

前記システムは更に前記第 1 の記憶装置システムと前記第 2 の記憶装置システムとに接続される第 4 の記憶装置システムを有し、

ジャーナル書き込みステップは、前記第 1 の記憶装置システムが時刻情報を有するジャーナルを前記第 3 の記憶装置システム若しくは前記第 4 の記憶装置システムのいずれかに書き込むステップを有し、

前記第 2 のコントローラはジャーナルを前記第 3 の記憶装置システム及び前記第 4 の記憶装置システムから取得して、取得したジャーナルに含まれるライトデータを、該ジャーナルに付与された時刻情報が示す時刻順に前記第 2 のディスク装置に書き込むことを特徴とするリモートコピー方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明はコンピュータによって使用されるデータを格納する記憶装置に関する。特に、コンピュータとの間でデータの授受を行う制御部と、データを格納するディスク装置とを内蔵する記憶サブシステムと、この記憶サブシステムから距離的に離れた場所に存在する他の記憶サブシステムとを相互に接続し、コンピュータを経由せずに記憶サブシステム間でデータをコピーして、データを二重化するリモートコピー技術に関する。

【 0 0 0 2 】

【従来の技術】

大型計算機システム、サーバー、ネットワーク上のパーソナルコンピュータ、又はその他の上位計算機システム（以下、ホストという。）からデータを受け取った記憶装置（以下、記憶サブシステムという。）が、遠隔地に設置された第2の記憶サブシステムに対して、データの順序性を保証しながら受信したデータを非同期転送し、第2の記憶サブシステムが転送されたデータを書き込む、非同期リモートコピーの技術がある。ここで非同期転送とは、記憶サブシステムが、ホストからデータを受信し、この受信データに対する処理の完了を通知する応答をホストに返信した後に、第2記憶サブシステムへのデータ転送を実行することを意味する。

【 0 0 0 3 】

また、ホストとこれに接続された第1記憶サブシステム間のデータ更新処理と同期させて、第1記憶サブシステムと第1記憶装置システムの付近地又は遠隔地に設置された第2の記憶サブシステムとの間でデータを同期転送する同期リモートコピー技術もある。ここで同期転送とは、記憶サブシステムが、ホストからデータを受信し、受信データを第2記憶サブシステムに転送した後に、ホストに対して応答を返信することを意味する。同期リモートコピー技術を用いれば、巨視的にみて2つの記憶サブシステムに格納されているデータが一致しており、データの書き込み順序も保証されている。尚、適当なデータ転送経路を選択すれば、2つの記憶サブシステムの距離が1 0 0 k mを超える場合であっても、同期転送によるコピーが可能である。

【 0 0 0 4 】

更に、3以上の記憶サブシステム間で、同期リモートコピーと非同期リモートコピーを組み合わせ、データ更新の順序性を保証したデータの二重化を実現する技術が特許文献1及び特許文献2に開示されている。

【0 0 0 5】

【特許文献1】

特開 2 0 0 0 - 3 0 5 8 5 6 号公報

【特許文献2】

特開 2 0 0 3 - 1 2 2 5 0 9 号公報

【0 0 0 6】

【発明が解決しようとする課題】

従来技術では、2箇所のデータ格納拠点（サイト）間でのリモートコピーを複数組み合わせ、nサイト間でリモートコピーを実行する。

【0 0 0 7】

従って、各サイトにデータのコピーを保持する必要があるため、各サイトはリモートコピー処理を実行しない場合と比べて最低でも n 倍の記憶容量を要し、コストが非常に高くなる。

【0 0 0 8】

またすべてのサイトにリモートコピープログラムを実装した記憶サブシステムが必要なため、高機能かつ高価な記憶サブシステムが複数台必要になる。

【0 0 0 9】

nサイトに渡る複数のリモートコピーペアの状態を監視・制御するため管理・制御が複雑となり開発コストも高くなる。

【0 0 1 0】

また、一般にリモートコピーに係わる処理の負荷は大きいため、正サイトと副サイトの間に位置し、正サイトとも副サイトともリモートコピー処理を実行する必要のある中間サイトの負荷は特に大きくなり、中間サイトの記憶サブシステムが処理可能なI/O数が制限されてしまう。

【0 0 1 1】

従って、より適切なnサイト間でのリモートコピー技術が求められている。

【0012】

【課題を解決するための手段】

第1の記憶装置システムと第2の記憶装置システムとを第3の記憶装置システムを介して接続する。リモートコピー処理を実行する際、第1の記憶装置システムは計算機から受信したライト要求に応じて、計算機から受信したライトデータとライトデータが書き込まれる格納位置を示すアドレス情報とを有するジャーナルを第3の記憶装置システムに送信し、第3の記憶装置にライトする。第2の記憶装置システムは、第1の記憶装置システムによって発行される制御情報を受信し、制御情報に基づいて第3の記憶装置からジャーナルをリードして取得する。そして第2の記憶装置システムは、ジャーナルに含まれるアドレス情報に従って、ジャーナルに含まれるライトデータを、第2の記憶装置システム内のディスクに書き込む。

【0013】

【発明の実施の形態】

ー第1の実施形態ー

図1に、本発明の第1の実施形態にかかるデータ処理システム（以下、リモートコピーシステムともいう。）の構成例を示す。なお、以下の説明では、正／副／中間を符号のa, b, cにより区別する。又、正／副／中間を区別しなくても良いときには符号のa, b, cを省く場合もある。

【0014】

データ処理システムは、正ホストコンピュータ（以下、正ホスト若しくは第1ホストと呼ぶ。）102a及び正記憶サブシステム104a（以下、第1記憶サブシステムと呼ぶ場合もある。）を有する正サイト（以下、第1サイトと呼ぶ場合もある。）101a、副ホスト（以下、第2ホストと呼ぶ場合もある。）102b及び副記憶サブシステム（以下、第2記憶サブシステムと呼ぶ場合もある。）104bを有する副サイト（以下、第2サイトと呼ぶ場合もある。）101b、中間記憶サブシステム（以下、第3記憶サブシステムと呼ぶ場合もある。）104cを有する中間サイト（以下、第3サイトと呼ぶ場合もある。）101cを有する。

【0015】

本実施形態では、記憶サブシステム104としてディスクアレイ装置を用いる例を説明するが、記憶サブシステム104は特にディスクアレイ装置に限られることはない。また、各サイト101内のホスト102、記憶サブシステム104の数に関しても、一以上であれば特に制限はない。また、各ホスト102は、ワークステーション、マイクロコンピュータ又はメインフレームコンピュータ等である。

【0016】

記憶サブシステム104は一つ以上のインタフェース110を持ちストレージエリアネットワーク（SAN103）を経由しホスト102と接続される。ここでインタフェース110はSCSI、ファイバチャネル、FICON、又はESCON等の記憶装置向けインタフェースであり、インタフェース110の種類およびSAN103の構成に関しては特に制限はない。以下本実施形態の説明においてはインタフェース110としてファイバチャネルを用いた例を説明する。

【0017】

異なるサイト101に存在するの記憶サブシステム104間も同様にインタフェース110（異なるサイト101に存在する記憶サブシステム104間の接続に用いられるインタフェースを以下、リモートコピーリンク116とも呼ぶ。）を介して接続される。

【0018】

正記憶サブシステム104aと中間記憶サブシステム104cは同じ部屋の中、または同じビルの中にあっても良いが、安全性のために（両装置が同時に同じ障害にあわないよう）距離をおいても良い。正記憶サブシステム104aと中間記憶サブシステム104cの間では同期リモートコピーが実行され、データの更新は同期で行われるため、正記憶サブシステム104aと中間記憶サブシステム104cの距離は短い方がよい。しかし、当該距離はシステムの立地条件等の環境により決定されるべきであり、特に制限はない。尚、通常通信の伝播遅延は1kmあたり5usといわれているので、当該距離が100km程度までであれば、ホスト102aへ著しい影響を与えることは少ないといわれている。

【0019】

副記憶サブシステム 104b は、安全性向上のために、正記憶サブシステム 104a や副記憶サブシステム 104b とは距離の離れた遠隔地に設置することが好ましい。

【0020】

尚記憶サブシステム 104 間の距離が長く、ファイバチャネルのデータ転送可能距離を越える場合には、ファイバチャネルに加え、エクステンダ装置を介して ATM、SONET、DWDM、IP などの広域回線を経由して、記憶サブシステム 104 間を接続しても良い。

【0021】

管理コンソール 117 は記憶サブシステム 104 を監視・管理するためのソフトウェアであり、CPU 202 やメモリ 205 を実装したホスト 102 上に実装される。管理コンソール 117 は災害時のことを考え、第四サイト（管理センタ）101d 上に設置され、第四サイト内のホスト上で稼動する。

【0022】

管理コンソール 117 は管理ネットワーク 111 を介して各記憶サブシステム 104 と接続され、記憶サブシステム 104 の状態の監視したり、データ処理システムの構成変更を実施したり、リモートコピー処理を制御したりする。一般に管理ネットワーク 111 は IP プロトコルを用いたネットワークにて構成されるが、他のプロトコルを用いたネットワークであってもかまわない。また、本実施例ではリモートコピーリンク 116 は管理ネットワーク 111 とは別のネットワークとしているが、管理ネットワークとして IP ネットワーク等を用い、更に管理ネットワーク 111 をリモートコピーリンク 116 として使用しても良い。

【0023】

本実施形態においては、正記憶サブシステム 104a と副記憶サブシステム 104b がリモートコピーのペアを構成している。一つの正記憶サブシステム 104a 内の正ボリューム 109 と一つ以上の副記憶サブシステム 104b 内の副ボリューム 115 がリモートコピーペアとして関連付される。

【0024】

リモートコピーの制御は記憶サブシステム 104 上のプログラム 105 によって実現される。正記憶サブシステム 104 a と副記憶サブシステム 104 b のプログラム 105 は基本的に同等の機能を持つので、同一記憶サブシステム 104 内に正ボリュームと副ボリューム 115 が混在していても良いし、災害又はメンテナンス時には、正ボリューム 109 と副ボリューム 115 の切り替えを実行することもできる。

【0025】

リモートコピーを実現するために、正記憶サブシステム 104 a は、通常の記憶サブシステム 104 が有する構成要素のほか、リモートコピープログラム 105 a、構成情報 106 a、リモートコピーキュー 114 の制御情報 107 a、リモートコピーの更新情報 108、リモートコピーされるデータを有する。同様に副記憶サブシステム 104 b は、通常の記憶サブシステム 104 の構成要素のほか、リモートコピープログラム 105 b、構成情報 106 b、リモートコピーキュー 114 の制御情報 107 b を有する。

【0026】

構成情報 106 はリモートコピーシステムが有する正記憶サブシステム 104a、中間記憶サブシステム 104 c、副記憶サブシステム 104 b 各々の構成、製品名、アドレス情報に加え、中間記憶サブシステム 104 c 内に格納される更新要求（以下ジャーナルとも呼ばれる。）の格納先アドレス及びキュー制御情報の格納先アドレスを有する。構成情報 106 のオリジナル情報は管理コンソール 117 が保持し、中間記憶サブシステム 104 c を経由して、または管理ネットワーク 111 を経由して、管理コンソール 117 から各記憶サブシステム 104 へ構成情報 106 の複製が送信され、各記憶サブシステム内に保存される。

【0027】

正記憶サブシステム 104a と副記憶サブシステム 104b は構成情報 106 のコピーを持ち、それぞれ記憶サブシステム 104 内のリモートコピープログラム 105 が自記憶サブシステム 104 内に格納されている構成情報 106 を参照してリモートコピーを実行する。

【0028】

正記憶サブシステム 1 0 4 a 及び副記憶サブシステム 1 0 4 b は各々、構成情報 1 0 6 から、中間記憶サブシステム 1 0 4 c 上のリモートコピーキュー領域 1 1 4 に格納された更新要求の位置情報、制御情報領域 1 1 3 に格納されたりリモートコピーキュー領域 1 1 4 に対する制御情報の位置情報を取得する。そして、正記憶サブシステム 1 0 4 a 及び副記憶サブシステム 1 0 4 b は各々、取得した位置情報を用いて中間記憶サブシステム 1 0 4 c 内のリモートコピーキューや制御情報領域にリード若しくはライト要求を送信する。これによって、正記憶サブシステムや副記憶サブシステムは、制御情報 1 0 7 を更新したり参照したりし、リモートコピーの更新要求やデータやメッセージを中間記憶サブシステムに書き込んだり、中間記憶サブシステムから読み出したりし、これらの情報を正記憶サブシステムと副記憶サブシステム間で送受信する。

【 0 0 2 9 】

尚、データの整合性を保つために、中間記憶サブシステム 1 0 4 c 内の各領域（例えば制御情報領域 1 1 3 やリモートコピーキュー領域 1 1 4 内の各記憶領域）には、所有者となるべき記憶サブシステム 1 0 4 を決めておき、データの更新は唯一に決められている所有者である記憶サブシステム 1 0 4 から実施する。

【 0 0 3 0 】

中間記憶サブシステム 1 0 4 c が記憶領域の排他制御機能を有している場合は、当該機能を使用して制御情報格納領域 1 1 3 やリモートコピーキュー領域 1 1 4 等に格納されるデータの整合性を保つことが望ましい。代表的な排他機能としては S C S I の R e s e r v e コマンドが挙げられる。R e s e r v e コマンドを用いれば、ロジカルユニット (L U) 単位の排他制御が可能であるため、中間記憶サブシステム 1 0 4 c は各記憶サブシステム 1 0 4 に別々の L U を割り当て、排他制御処理を実施すれば良い。この際中間記憶サブシステム 1 0 4 c は、各領域の必要サイズに応じて L U サイズを設定することが望ましい。尚、本発明においては、排他制御の対象となる記憶領域の単位、排他制御の方式については特に制限はなく、例えば L U より小さな単位（たとえばブロック単位等）で排他制御が実行されてもかまわない。

【 0 0 3 1 】

リモートコピーキューの制御情報 1 0 7 にはリモートコピーキュー 1 1 4 を制御するための情報が含まれる。正記憶サブシステム 1 0 4 a と副記憶サブシステム 1 0 4 b は各々、自装置が有するリモートコピーキューの制御情報 1 0 7 を参照してリモートコピーキュー 1 1 4 に対し変更を行い、中間記憶サブシステム 1 0 4 c が有する制御情報 1 1 3 を更新することによって、変更後の状態をリモートコピーペアの相手方の記憶サブシステム 1 0 4 (正記憶サブシステムから副記憶サブシステムへ若しくは副記憶サブシステムから正記憶サブシステム) へ伝える。

【 0 0 3 2 】

正記憶サブシステム 1 0 4 a が有する更新情報 1 0 8 とは、正ホスト 1 0 2 a から正記憶サブシステム 1 0 4 a が受信するライト要求についての、リモートコピーに関する情報で、各ライトデータの書き込み時刻、書き込み位置、サイズが含まれる。更新情報 1 0 8 は正記憶サブシステム 1 0 4 a が正ホスト 1 0 2 a からライト要求を受け付けたときに正記憶サブシステム 1 0 4 a によって生成され、正記憶サブシステムが中間記憶サブシステム 1 0 4 c に送信する更新要求にも含まれるデータである。尚、更新情報 1 0 8 は正記憶サブシステム 1 0 4 a が発行したリモートコピーの更新要求が副記憶サブシステム 1 0 4 b によって取得されたことが確認された後に削除される。中間記憶サブシステム 1 0 4 c に障害が発生したときには更新情報 1 0 8 を使用し、正記憶サブシステム 1 0 4 a と副記憶サブシステム 1 0 4 c の同期を行う。

【 0 0 3 3 】

尚、ここまでで説明したプログラム 1 0 5 は、コンパクトディスクや光磁気ディスクといった可搬記憶媒体を用いて、あるいは、管理ネットワーク 1 1 1 を介して、他の装置から各記憶サブシステムが有する記憶媒体にインストールされ、各記憶サブシステムが有する CPU によって実行される。

【 0 0 3 4 】

図 2 に記憶サブシステム 1 0 4 の構成例を示す。

【 0 0 3 5 】

各記憶サブシステム 1 0 4 は、コントローラ 2 0 1、一つ以上のディスク装置

210を有する。コントローラ201は、ホストと接続するためのホストアダプタ203、メモリ205、ディスク装置と接続するためのディスクアダプタ207、プロセッサ202、管理系ネットワークと接続するためのネットワークコントローラ204を有する。本発明では、記憶サブシステムが有する各構成要素の数には特に制限はないが性能や信頼性の観点から各構成要素は多重化するほうが好ましい。

【0036】

ホストアダプタ203はファイバチャネル等のインタフェース110に関するプロトコル制御を行う。ディスクアダプタ207はファイバチャネル等のディスクディスクインタフェース209に関するプロトコル処理を行う。

【0037】

メモリ205上にはリモートコピー処理に用いられるデータやプログラムが格納されている。即ちメモリ205には、リモートコピーを実現するために、リモートコピーのプログラム105、リモートコピーの構成情報106、制御情報107、更新情報108、データ211が格納される。これらの情報以外にメモリ205上には記憶サブシステム104を制御するために必要なプログラムや制御情報、ホスト102から受信するデータ211も格納される。高信頼化のためにメモリ205の二重化やメモリ205への電源の二重化をしておくことが好ましい。

【0038】

ネットワークコントローラ204は管理系ネットワーク111のプロトコル制御を行い、管理コンソール117と記憶サブシステム104の間の通信を実施する。

【0039】

ディスクドライブ210はディスクインタフェース209を介してコントローラ201からのリードやライト等のコマンドを受付け、コマンドに従ってリード若しくはライト処理を行う。高信頼化のため、ディスクインタフェース209は二重化しておくことが好ましい。一般に記憶サブシステムにおいては、ディスクドライブ210を複数組み合わせる冗長構成をとり、複数のディスクドライブの

中に論理的なデバイス（以下、ボリュームともいう。）を作成する。以下、論理的なデバイスであるボリュームを用いて実施形態を説明する。

【0040】

プロセッサ202は記憶サブシステム104に関する処理を行う。プロセッサ202は内部バス208を介しコントローラ201内のホストアダプタ203、ディスクアダプタ207、ネットワークコントローラ204と接続しこれらを制御する。更にプロセッサ202はメモリ205とも内部バス208を介し接続し、メモリ205内の構成情報106や制御情報107を用いながらプログラム105を実行したり、制御情報107を更新したりする。

【0041】

本実施例においては簡素な内部構成を有する記憶サブシステムの例を説明したが、上述の説明と同等の機能が実現可能であれば、記憶サブシステム104の内部構成に関しては特に制約はない。例えば特開平10-333836に開示されている、内部バス208の代わりにスイッチを使用した記憶サブシステムを用いても良い。

【0042】

図3に中間記憶サブシステム104cのディスクドライブ内に配置される記憶領域の構成例を示す。中間記憶サブシステム104cのディスクドライブ内には、構成情報106が格納される構成情報領域112、制御情報107が格納される制御情報領域113、更新要求が格納されるリモートコピーキュー領域114の3つの領域がある。

【0043】

構成情報106には、リモートコピーシステムが有する正記憶サブシステム、中間記憶サブシステム、及び副記憶サブシステム各々の構成、製品名、アドレス情報に加え、リモートコピーキュー領域114内の更新要求の格納位置や、制御情報領域113内のキュー制御情報107の格納先アドレスが含まれる。

【0044】

制御情報107には、リモートコピーキュー114を制御・管理するための情報が含まれる。正記憶サブシステム104aから副記憶サブシステム104bに

宛てた（正→副）制御情報 107 を格納するための正制御情報領域 301 と、副記憶サブシステム 104b から正記憶サブシステム 104a に宛てた（副→正）制御情報 107 を格納するための副制御情報領域 302 の二種類の領域をもち、正制御情報領域 301 は正記憶サブシステム 104a が、副制御情報領域 302 は副記憶サブシステム 104b が所有権を持つことによって、データの不整合発生を防ぐ。尚、各記憶サブシステムが有する構成情報 106 には、正制御情報領域 301、副制御情報領域 302 等の各領域のアドレスやサイズ以外に、各領域の所有者を示す情報が含まれており、リモートコピープログラム 105 が構成情報を参照して所有者を確認することで、データの不整合が発生しないように制御を行う。SCSI の Reserve コマンドのように物理的な排他が可能な場合はより堅牢なシステムが構成可能になる。この場合には、中間記憶サブシステム 104c が論理ユニット（LU）ごとに正制御情報領域、副制御情報領域を割り当て、LU ごとに排他制御を行うことになる。

【0045】

リモートコピーキュー 114 に格納される情報には、大別してメッセージとデータの 2 種類の情報があるため、各情報を格納するためにリモートコピーキュー 114 にはメッセージ領域とデータ領域の 2 種類の領域がある。以下、2 種類の領域を特に区別しない場合はキュー 114 として説明する。キュー 114 には、制御情報 107 と同様、正記憶サブシステム 104a から副記憶サブシステム 104b に宛てた（正→副）情報と副記憶サブシステム 104b から正記憶サブシステム 104a へ宛てた情報の両方が格納される。従って、キュー 114 には少なくとも、正記憶サブシステム 104a から副記憶サブシステム 104b に宛てて発行される、メッセージを格納する正メッセージ領域 303 とデータを格納する正データ領域 305、副記憶サブシステム 104b から正記憶サブシステム 104a に宛てて発行される、メッセージを格納する副メッセージ領域 304 とデータを格納する副データ領域 306 の 4 つの領域が存在する。尚、リモートコピーキュー 114 に存在する記憶領域の数はリモートコピーシステムの構成に依存し、例えば副記憶サブシステムを複数有するリモートコピーシステムであれば、リモートコピーキュー 114 に存在する記憶領域の数は 4 より多くなることもある。

【 0 0 4 6 】

制御情報領域 1 1 3 と同様キュー領域 1 1 4 においても、正記憶サブシステム 1 0 4 a から副記憶サブシステム 1 0 4 b に宛てた情報（メッセージ若しくはデータ）が格納される領域は正記憶サブシステムが所有者となっており、副記憶サブシステム 104b から正記憶サブシステム 104a に宛てた情報は副サブシステムが所有者となっている。

【 0 0 4 7 】

図 4 に中間記憶サブシステム 1 0 4 c のディスクドライブに格納されるデータの構造例を示す。

【 0 0 4 8 】

中間記憶サブシステム 1 0 4 c のディスクドライブ 2 1 0 は、通常のディスクドライブであり、中間記憶サブシステム 1 0 4 c も通常の記憶サブシステムであるため、ディスクドライブ 2 1 0 に書き込まれた情報は他の装置から上書きすることができる。従ってディスクドライブ 2 1 0 内の記憶領域に書き込まれたデータが、所有者の論理的障害（プログラムのバグ）または、当該記憶領域の所有者以外の他の装置によって上書きされ、データがエラーになる（不整合となる）可能性がある。中間記憶サブシステム 1 0 4 c でこれらのエラーを検知可能であればエラーに対する対応が可能であるが、所有者以外のホスト 1 0 2 又は記憶サブシステム 1 0 4 からのデータ上書き、論理エラー等を中間記憶サブシステムが検知するのは困難な場合もある。

【 0 0 4 9 】

そこで本実施形態においては、中間記憶サブシステム 1 0 4 c に書き込まれるデータに保護データを付加することにより、正記憶サブシステム 1 0 4 a と副記憶サブシステム 1 0 4 b の間で中間記憶サブシステム 1 0 4 c を介して授受されるデータの保障を行う。

【 0 0 5 0 】

保護データが付加されたデータの例を図 4 に示す。尚本実施形態においては、中間記憶サブシステム 1 0 4 c は S C S I 等のブロックデバイスを有する記憶サブシステムであると仮定している。通常ブロックは 5 1 2 バイトのサイズを有し

、中間記憶サブシステム 104c が有するディスクデバイスは、ブロック単位でアクセスされる。

【0051】

図4(a)は、中間記憶サブシステム 104c が有するディスクデバイスの各ブロック 401 の最後の領域を保護データ領域 402 とした例である。保護データ 402 を各ブロックの最後には書き込むため、従来はブロックの最後の領域に格納することのできた元データが次のブロックの先頭位置 (403) に格納されることになる。従来次のブロックの先頭であった位置 (404) は記憶領域 403 の後になるため、この方法を用いるとデータの格納位置が順次後ろにずれていく。

【0052】

このため従来一ブロック内に格納されていたデータが、保護データ領域を設けたことにより複数ブロックにまたがって格納されることになる。すると、一ブロック分のデータを中間記憶サブシステム 104 に格納するために複数ブロックを更新する必要が生ずる。ブロックデバイスはブロック単位でアクセスされるため、この場合は関連するブロックをすべてディスクデバイスからコントローラ内のメモリ 205 上に読み込み、データを更新し、これをディスクデバイス 210 に書き込む必要がある。しかし、本発明において中間記憶サブシステム 104c に格納されるデータは、一度書き込まれるのみで、同一データの読み書きが行われないため図4(a)に示す方法を採用しても問題は少ない。

【0053】

図4(b)は中間記憶サブシステム 104c のブロックサイズを保護データ分だけ増加させ、増やした領域に保護データ 402 を格納する例である。この方法を用いるとブロックアドレスを変更することなく、保護データ 402 が追加されたデータをディスクデバイスに格納することが可能である。

【0054】

図5に保護データ 402 の一例を示す。図4において説明したとおり保護データ 402 は中間記憶サブシステム 104c に格納される各ブロックに付加されている。保護データ 402 には大別して2種類のデータが含まれる。一つは要求の

識別子 501 と要求内の通し番号（シーケンス番号） 502 とを有する論理的な情報である。他の一つはブロックデータのエラーをチェックするための誤り検出符号 503 である。前者は論理的な誤りや他者によるデータ改ざんを検出するために用いられ、後者はブロック自体のデータ障害を検出するために用いられる。誤り検出符号 503 にはパリティ、チェックサム、ECC、CRC、ハミング符号等があるが本発明においては誤り検出及び実装が可能であればどの方式の誤り検出符号を用いてもかまわない。

【0055】

保護データ 402 はデータを生成する記憶サブシステム 104 のホストアダプタ 203 にてデータに付加され、受信側の記憶サブシステム 104 のホストアダプタ 203 で保護データを用いたデータのチェックが行われる。

【0056】

図 6 に中間記憶サブシステム 104c に格納されるデータの論理構造の一例を示す。図 6 は、あるリモートコピーグループ内の正記憶サブシステム 104a から副記憶サブシステム 104b へ送信される更新要求に関する情報を示している。

【0057】

構成情報領域 112 内に格納されている構成情報 106 には静的な構成の情報が含まれ、構成情報 106 はリモートコピーグループの構成変更時に更新される。各記憶サブシステム 104 は構成情報 106 を中間記憶サブシステム 104c の構成情報格納領域 112 から取得して自装置内に展開し、構成情報からリモートコピーに関する情報およびリモートコピーキュー領域 114 に関する構成情報を取得する。そして、これらの情報を利用して各記憶装置サブシステムはリモートコピーおよびキュー制御を実行する。

【0058】

本実施例では中間記憶サブシステム 104c を介して各記憶サブシステム 104 へ構成情報 106 を配布しているが、構成情報 106 は管理コンソール 117 から各記憶サブシステム 104 へ直接送付しても良い。

【0059】

図6では一つのリモートコピーグループに関する構成情報106を示している。

【0060】

通常リモートコピーグループとは一つ以上のボリュームから構成され、各ボリューム間の整合性（コンシステンシ）を保つグループとして定義される。リモートコピーグループに対するサスペンド、再同期等の操作は、ボリュームペア単位その他、グループ単位でも実行可能である。リモートコピーグループは一つの正記憶サブシステム104aと一つ以上の副記憶サブシステム104bで構成され、識別子（グループID）601が付けられている。従って構成情報106には、グループID601、リモートコピーグループに属する正記憶サブシステムのID602と副記憶サブシステムのID604、副記憶サブシステム104bの個数603が含まれる。さらに構成情報106には、キュー領域114内に存在する正メッセージ領域や正データ領域等の領域（以下、キューとも呼ぶ）の数605と各キューの個別情報606が格納される。

【0061】

図3に示したように、正記憶サブシステム104と副記憶サブシステム104が一对一の基本構成では、メッセージとデータに関して双方向のキューが生成されるため、一つのリモートコピーグループに関して4つのキューが存在する。図6では正記憶サブシステム104aから副記憶サブシステム104bへ宛てて送信される正データのキューに関する個別情報を示している。逆方向のキューに関しても、正・副が入れ替わるだけで、同様の制御が実施される。

【0062】

キューの個別情報の中にはキューの識別子607、メッセージ又はデータのいずれに関するキューかを識別するためのキュー種別608、中間記憶サブシステム104cのキュー領域114内における当該キューの先頭位置609、キューのサイズ610、正記憶サブシステム104aから副記憶サブシステム104bへ送付される制御情報107に関する正情報611、副記憶サブシステム104bから正記憶サブシステム104aへ送付される制御情報107に関する副情報612が含まれる。

【0063】

尚、構成情報領域 112 とキュー領域 114 とが別々の記憶サブシステム 104 に存在する場合には、先頭位置 609 として、キュー領域が配置される記憶サブシステムの識別子、当該キューの LU 等の論理アドレス等が含まれる。副記憶サブシステムが複数個存在する場合にはこれらの情報も複数個必要になる。

【0064】

制御情報 107 に関する情報 611 や 612 には、制御情報の発行元記憶サブシステム 104 の識別子 613、制御情報領域 113 内での制御情報の格納位置を示す先頭位置 614、制御情報のサイズ 615 が含まれる。各記憶サブシステム 104 は構成情報 106 内の制御情報に関する情報（例えば 611 や 612）を元に、制御情報 107 が格納されている記憶領域を特定して制御情報 107 をやり取りする。

【0065】

一つのキューに対応する制御情報 107 としては、一つの正制御情報 631 と一つ以上の副制御情報 632 がある。各制御情報 107 は予め定められた唯一の記憶サブシステム 104 によって更新されるが、制御情報の参照は複数の記憶サブシステム 104 から可能である。

【0066】

各制御情報 107 は、リモートコピーのグループ識別子 616、キューの識別子 617、当該制御情報 107 の所有装置を示す所有装置識別子 618、リモートコピー処理の進捗状況を示すためリモートコピー中データの先頭位置を示す先頭位置 619、リモートコピー中のデータのサイズ 620、及びハートビート 621 を有する。

【0067】

グループ ID 616、キュー ID 617、所有装置識別子 618 は、制御領域が論理的に正しいかをチェックするために使用される。先頭位置 619 及びサイズ 620 には各々、リモートコピー処理中のデータが格納されているキュー領域内の記憶領域の先頭位置と、当該記憶領域のサイズが格納される。更に各記憶サブシステム 104 はハートビート 621 に定期的に情報を書き込むことで、自装

置が稼動中であることを他の装置に通知する。言い換えれば、ハートビート621が更新されているか否かを確認することで、当該ハートビートに対応する記憶サブシステムが稼動中であるか否かを他の装置から判別できる。尚、ハートビート領域621に書き込まれる情報はカウンタ値やタイマなど時間に伴い変化する値であることが望ましい。

【0068】

正記憶サブシステム104aと副記憶サブシステム104bの制御、動作は同期していないため、中間記憶サブシステム104cの制御情報領域113に格納されている正制御情報631の先頭位置619及びサイズ620によって示されるデータと、副制御情報632の先頭位置619及びサイズ620によって示されるデータと、正記憶サブシステム104aが有する制御情報107aの先頭位置626及びサイズ627によって示されるデータと、副記憶サブシステム104bが有する制御情報107bの先頭位置及びサイズによって示されるデータとは同一とは限らない。

【0069】

図6を用いて正制御情報631、副制御情報632、正記憶サブシステム104aが有する制御情報107a、及び副記憶サブシステム104bが有する制御情報107bの関連を説明する。

【0070】

キュー領域114にはData1～Data6が格納されているものとする。また、図6において中間記憶サブシステム104c上の正制御情報631は、先頭位置619がData1を示し、サイズ620がData1～Data4までのサイズを示しており、この結果Data1を先頭にData4までを処理中の要求（処理中のデータ）として示しているものとする。副制御情報632は、先頭位置619がData3を示し、サイズ620がData3分のサイズを示しており、この結果Data3を処理中の要求（処理中のデータ）として示しているものとする。

【0071】

更に、正記憶サブシステム104aが有する制御情報107aは先頭位置626がData1を示し、サイズがData1からData6までのデータのサイズを示しているものとする。即ち結局制御情報107aは、Data1からData6までを示してい

ることとなり、正記憶サブシステム104aはData1～Data6をリモートコピー中のデータとして認識しているものとする。正記憶サブシステム104aが次のデータをキュー領域114に書き込む場合には、制御情報107aの次位置628が示す、Data6の次の位置からデータを書き込む。

【0072】

一方副記憶サブシステム104bはData1、及びData2をすでに中間記憶サブシステム104cから取得しており、副記憶サブシステム104bが有する制御情報107bの先頭位置629はData3を、サイズ630はData3分のデータ量を示しているものとする。従って、副記憶サブシステム104bは、Data3を処理中の要求（処理中のデータ）と認識している。

【0073】

尚、副記憶サブシステム104bは、中間記憶サブシステム104c上の正制御情報631を参照することによって、正記憶サブシステムがData1を先頭624にData4までを処理中の要求（処理中のデータ）として認識している、と判断する。従って、副記憶サブシステム104bは、既に処理したData1、及びData2と、処理中のData3のみならず、Data4も処理すべきことを、中間記憶サブシステム104cに格納されている正制御情報631を参照することによって、認識することができる。

【0074】

Data5及びData6は、中間記憶サブシステム104c上の正制御情報631が更新された後に正記憶サブシステム104aからキュー領域114に追加された情報である。キュー領域へのデータの書き込みと、正制御情報631の更新とは非同期で実行されるので、図6においては、Data5及びData6のキュー領域114への追加に関しては正制御情報631にはまだ更新されていない。最新の制御情報は正記憶サブシステム104a内に制御情報107aとして存在し、一定時間後、正記憶サブシステム104aから中間記憶サブシステム104cへの正制御情報631更新契機が訪れた際に、正制御情報631が最新情報に更新される。

【0075】

次に、正記憶サブシステム104aが中間記憶サブシステム104c上の正制御情

報631やキュー領域内のデータ（以下、ジャーナル若しくは更新要求ともいう）、又は正記憶サブシステムが有する制御情報107aを更新する処理について説明する。

【0076】

正制御情報631及びキュー領域内のデータは、正記憶サブシステム104aが、ホスト102からのライト要求と同期または非同期にて更新する。

【0077】

正記憶サブシステム104aが有する制御情報107aの次位置628は、正記憶サブシステム104aがキュー領域114へ更新要求を送付するのと同期して、正記憶サブシステム104a内で更新される。正記憶サブシステム104aが有する制御情報107aの先頭位置626は、中間記憶サブシステム104cが有する副制御情報632の先頭位置619に基づいて、正記憶サブシステム104aによって、ホストからのライト要求とは非同期にて更新され、トラフィックを削減するために複数ライト分を更新可能である。正記憶サブシステム104aの制御情報107aのサイズ627に関しては、先頭位置626と次位置628各々の更新に合わせて更新される。

【0078】

正記憶サブシステム104aは定期的に中間記憶サブシステム104cの制御情報領域113に格納されたデータにアクセスする。このとき正記憶サブシステム104aは、副制御情報632を取得し（リードし）、副制御情報632の先頭位置619を参照して、正記憶サブシステムが有する制御情報107aの先頭位置626を変更する。例えば上述の例においては、制御情報107aの先頭位置626は現在Data1を示しているが、副制御情報632の先頭位置619はData3を示しているため、正記憶サブシステムは中間記憶サブシステム104cの制御情報領域113にアクセスした際に、副制御情報632を取得して、自装置の制御情報107aの先頭位置626をData3に更新する。

【0079】

この際正記憶サブシステム104aに格納されている更新情報108もあわせて削除可能となる。即ち、正記憶サブシステムは副制御情報632を参照することによって、Data1及びData2については副記憶サブシステムによって取得済みで

あることが認識できるので、Data 1 及びData 2 についての更新情報 1 0 8 は削除できる。削除のタイミングは、先頭位置626の更新時点以降であればどのタイミングでもよい。

【 0 0 8 0 】

また、正記憶サブシステム104aは、制御情報107aの正サイズ627も、正記憶サブシステム104aの制御情報107aが有する先頭位置626と次位置628から算出して更新する。係る処理によって制御情報107aの正サイズは、Data3からData6までのデータ量を示すことになる。

【 0 0 8 1 】

以上の処理により、正記憶サブシステムが有する制御情報107aは最新のデータに更新されるので、正記憶サブシステムはこれを中間記憶サブシステムの制御情報領域113に書き込むことにより、正制御情報631を更新する。

【 0 0 8 2 】

次に副記憶サブシステム 1 0 4 b が中間記憶サブシステム 1 0 4 c の制御情報領域113にアクセスし、副制御情報 6 3 2 を更新する処理について説明する。

【 0 0 8 3 】

副記憶サブシステム 1 0 4 b は一定時間間隔で中間記憶サブシステム 1 0 4 c から正制御情報 6 3 1 を取得し、正制御情報631が示す更新要求のうち、未だ取得していない更新要求をキュー領域 1 1 4 から取得する。その後、副記憶サブシステム104bは、自装置が有する制御情報107bと副制御情報 6 3 2 を更新する。

【 0 0 8 4 】

上述の例においては、最初正制御情報 6 3 1 はD a t a 1 からD a t a 4 までを処理中のデータとして示し、副制御情報 6 3 2 の先頭位置619はD a t a 3 を示している。従って、副記憶サブシステム104bは正制御情報631を参照することによって、Data 3 とData4を自装置が未だ取得していないことを認識する。そこで副制御サブシステムはD a t a 3 とD a t a 4 をキュー領域114から取得した後、自装置 1 0 4 b が有する制御情報107b及び中間記憶サブシステム 1 0 4 c が有する副制御情報632を更新する。即ち、制御情報107bの先頭位置629と副制御情報632の先頭位置619は共にData4を、制御情報107bのサイズ630と副制御情報632

のサイズ620は共に0を示すよう、制御情報が更新される。

【0 0 8 5】

尚、副記憶サブシステム 1 0 4 bの稼働率が高い等、何らかの理由で副記憶サブシステム 1 0 4 bがキュー領域114から更新要求を取得しなかった場合には、副記憶サブシステム 1 0 4 bは制御情報の更新のみを行う。この場合、制御情報107bの先頭位置629と副制御情報632の先頭位置619は共にData3を、制御情報107bのサイズ630と副制御情報632のサイズ620は共にData3とData4を合計したデータ量を示すよう、制御情報が更新される。

【0 0 8 6】

図 7 に、正記憶サブシステム若しくは副記憶サブシステムから発行され、キュー内に格納される各要求 7 2 1 のデータ構造例を示す。

【0 0 8 7】

要求の先頭にはヘッダ情報 7 2 2 が格納され、末尾には末尾情報 7 2 4 が格納される。ヘッダ情報 7 2 1 には要求の属性 7 0 1、グループ I D 7 0 2、キュー I D 7 0 3 等の要求の属性と、要求の個別情報が格納される。要求の属性は論理エラーチェックのために用いられる。要求の個別情報には時系列情報と位置情報の二種類の情報がある。時系列情報としてはキュー内における要求の通番（要求 I D）7 0 4、正記憶サブシステム 1 0 4 a またはホスト 1 0 2 にて付加されたタイムスタンプ 7 0 5 があり、副記憶サブシステム 1 0 4 b 内で要求を時系列順に整列し、要求の抜けをチェックするときに用いられる。位置情報にはボリュームの I D 7 0 6、ボリューム内アドレス 7 0 7、要求のサイズ 7 0 8 があり、これらの位置情報を元に要求に含まれるデータが副記憶サブシステム 1 0 4 b 内に格納される。尚、要求に含まれる位置情報は、ホストから受信するライト要求に含まれる位置情報と同じものである。サイズ 7 0 8 はホストから受信するライトデータのサイズを示し、このサイズ 7 0 8 にヘッダ情報 7 2 1 と末尾情報 7 2 4 の各固定サイズを足したサイズが要求のサイズとなる。

【0 0 8 8】

データ 7 2 3 はホストから受信するライトデータである。

【0 0 8 9】

末尾情報 724 にはヘッダ情報 722 に加え誤り検出符号 709 が含まれる。誤り検出符号 709 はヘッダ情報 722 とデータ 633 とを通して計算され、要求全体に関する誤りの検出を行う。要求の誤り検出符号 709 は図 5 で示した保護データ 402 と併用することでより高信頼なリモートコピー処理を実現する。

【0090】

このように、ヘッダ情報 722 や末尾情報 724 等の制御情報とライトデータ 723 を連続した領域に書き込むことで、ホストからライト要求を受信した際にこれに対応して中間記憶サブシステム 104c へ発行される要求は、連続領域に対するライト要求となり、中間記憶サブシステム 104c に対する書き込みは一度行うだけでよい。

【0091】

尚、上述のように各要求にはホスト 102 から受信する位置情報等の制御情報とライトデータとが双方含まれており、ジャーナルとも呼ばれる。

【0092】

図 8 は本実施形態における初期設定の手順の一例を示すフローチャートである。最初に管理者は管理コンソール 117 を介して、リモートコピーシステムを構成する装置の情報、各記憶サブシステム内のボリューム等の記憶サブシステム 104 の情報、リモートコピーシステム内に存在するホスト 102 上で稼動するアプリケーションの情報や、ホスト 102 が使用しているボリュームの情報を取得する（801）。

【0093】

次に管理者は、管理コンソールが収集したこれらの情報を元に、リモートコピーのボリュームペアや、ボリュームペアを集めたコンシステンシグループ等を決定し、各記憶サブシステム 104 のリモートコピーの構成情報 106 を作成して管理コンソール 117 に入力する（802）。構成情報 106 については、管理コンソール 117 がオリジナル情報を保持し、各記憶サブシステム 104 へは中間記憶サブシステム 104c を経由して、または管理ネットワーク 111 を経由してコピーが送信される（803）。

【0094】

各記憶サブシステム 104 に構成情報 106 が設定され、リモートコピーのペアが確立されたり、コンシステンシグループが生成された後、正記憶サブシステム 104 a から中間記憶サブシステム 104c を介して副記憶サブシステム 104 b へのデータの初期コピーが実行され、リモートコピー処理が開始される (804)。

【0095】

図 9 に、本実施形態において正記憶サブシステム 104 a がホスト 102 からライト要求を受信した場合に実行する更新処理の一例を示す。

【0096】

正記憶サブシステム 104 a はホスト 102 からライト要求を受け付ける (901)。そして正記憶サブシステム 104a は、ライト要求から関連するボリューム、アドレスを算出し、以下に示すチェック処理を行う。まず正記憶サブシステムは、ライト要求の対象であるボリュームに関して、リモートコピー属性が指定されているかどうかを調べ、リモートコピー属性が指定されていない場合は、更新データ (以下ライトデータともいう。) をホスト 102 から受領し (911)、ホスト 102 へライト要求に対する処理の終了を報告して (909) 処理を終える。

【0097】

リモートコピー属性が指定されている場合は、ステップ 903 へ進む。正記憶サブシステム 104a は、ライト対象ボリュームに対応するリモートコピーペアの状態をチェックし、正常にペアが構成されている状態 (ペア状態) か、ペアが切り離されている状態か (サスペンド) をチェックする (903)。

【0098】

サスペンド状態になっている場合には差分情報を記録し (910)、更新データをホスト 102 から受領して (911)、ホスト 102 へ更新処理の終了を報告し (909)、処理を終える。尚差分情報は、任意のサイズの位置情報を一ビットに対応させたビットマップの形式で保持される。ビットマップで差分情報を保存する場合には、同一アドレスへの上書きは同じビットで示されるため、差分情報を格納するために確保しなければならない領域が少なくすむ反面、ホスト

から受信するライト要求の順序性（受信順序）を保存することはできない。順序性を保持したい場合には更新ログ等、各ライト要求の受信時刻、ライト対象記憶領域のアドレス、ライトデータのサイズ、ライトデータを保持する必要がある、ビットマップで差分情報を管理する場合より多くの記憶容量が必要になる。通常サスペンドは長時間にわたることが多いためビットマップで差分情報を保持することが多い。ビットマップは記憶サブシステム104のメモリ205上に格納してもよいが、ディスク210上に格納してもよい。

【0099】

ペア状態である場合には、正記憶サブシステム104aは中間記憶サブシステム104c上の記憶領域の有無をチェックする。具体的には正記憶サブシステム104aが保持する構成情報106aと制御情報107aから中間記憶サブシステム104c内に存在するキュー領域114の残りの記憶容量を計算する（904）。

【0100】

通常中間記憶サブシステム104cには、ホスト102の負荷で、即ちホストからのライト要求及びライトデータによってキュー領域114が全て埋ってしまうことのないよう、十分大きなキュー領域114を用意しておく。ただし予期しない負荷が発生する場合もあり、この場合は中間記憶サブシステム104cのキュー領域を使い切ってしまう。

【0101】

図9では中間記憶サブシステム104cのキュー領域に空き容量がない場合には、正記憶サブシステム104aがホスト102へビジー報告を返し、ライト要求を受け付けない（912）。ホスト102へビジーを返送する代わりに、正記憶サブシステム104aは処理（912）において中間記憶サブシステムに障害が発生していると判定し、リモートコピーの状態をサスペンド状態としてもよい。この場合には正記憶サブシステム104aは、図11に示す障害処理を実施し、リモートコピーの状態をサスペンドへ移行し、ホストから更新データ（ライトデータ）を受信してビットマップを更新し、ホストに対し通常の終了報告を行う。

【0102】

ステップ904にて中間記憶サブシステム104cに空き容量があると判断された

場合、即ちステップ902から904までの処理にてリモートコピー処理が可能であると判定された場合には、正記憶サブシステム104aはステップ905以下の処理を実施する。まず正記憶サブシステム104aはホスト102からライト要求に対応するライトデータを受領する（905）。

【0103】

次に正記憶サブシステム104aは、該ライト要求に対応するリモートコピーの更新要求を作成し、ライトデータと合わせて図7に示すような更新要求を作成してこの更新要求を中間記憶サブシステム104cのキュー領域114の次位置以降に書き込む（907）。このとき更新要求を書き込もうとするとキュー領域114の最末尾を超えてしまう場合は、超える分の続きのデータはキュー領域114の先頭から書き込む。尚、更新要求をキュー領域114に書き込む際には、副記憶サブシステム104bが有する制御情報107bが示す先頭位置629を越えないように正記憶サブシステム104aは制御する必要がある。

【0104】

次に正記憶サブシステム104aは、キュー領域114に書き込んだ更新要求のサイズ分だけ、制御情報107aの次位置628、サイズ627を変更し（908）、ホスト102へ終了を報告した後（909）処理を終える。

【0105】

図10に、正記憶サブシステム104aと中間記憶サブシステム104cとの間で実行される制御情報の更新処理の一例を示す。まず正記憶サブシステム104aは中間記憶サブシステム104cから、副記憶サブシステム104bの制御情報である副制御情報632を取得する（1001）。

【0106】

そして正記憶サブシステムは、副制御情報632内のハートビートを参照し副記憶サブシステム104bが稼動しているか否かをチェックする。具体的には、正記憶サブシステム104aは過去に取得した副記憶サブシステムのハートビート情報の内最も新しいハートビート情報を正記憶サブシステム104a内のメモリ205に保存しておき、中間記憶サブシステム104cから新たに副制御情報32を取得した際には、取得した副制御情報632中のハートビート情報と、メモリ

に格納されているハートビート情報とを比較して、チェックを行う。2つのハートビート情報が同じ値であれば副記憶サブシステムは稼動していないこととなる。

【0107】

尚、制御情報の更新処理は正記憶サブシステム104aと副記憶サブシステム104bとの間では非同期に実施されるため、一度ハートビート情報が更新されていないと判断されただけで副記憶サブシステム104bが稼動していないと判断するのではなく、数回連続してハートビート情報の更新がない場合や一定時間以上更新がない場合等に幅を広げて、副記憶サブシステム104bの稼動状況を判断する。たとえば、副制御情報632の更新周期が1秒である場合は、5秒以上ハートビート情報の更新がない場合に副記憶サブシステムが稼動していないと判断する等とする。

【0108】

ステップ1002において副記憶サブシステムが稼動していないと判断された場合には、正記憶サブシステム104aが図11に示す障害時処理を行う（1006）。

【0109】

ステップ1002において副記憶サブシステムが稼動していると判断された場合には、正記憶サブシステム104aは中間記憶サブシステム104cの制御情報領域113に格納されている副制御情報632を参照して、自装置が有する制御情報107aを更新する。即ち、制御情報107aの正先頭位置626を副制御情報632の副先頭位置619に合わせ、制御情報107aの正サイズ627は更新後の正先頭位置626から正次位置628までのサイズとする。そして正記憶サブシステム104aの新制御情報107aを用いて、中間記憶サブシステム104cの制御情報領域113に格納されている正制御情報631を更新する（1003）。

【0110】

次に正記憶サブシステムは更新情報108を廃棄する。副記憶サブシステム104bで新たに更新された更新要求、即ちステップ1003における変更前の先頭位置から変更後の先頭位置までに存在する更新要求は、既に副記憶サブシステム

によって取得されている。従ってこの更新要求に対応する更新情報については、正記憶制御装置にて保持する必要がないため、正記憶サブシステムはこの更新情報108を廃棄可能とし、任意の時間に破棄する（1004）。

【0111】

更に正記憶サブシステム104aは、ステップ1001からステップ1004の処理を一定間隔で実施するため一定時間（例えば1秒間）待機した後（1005）、ステップ1001から再び処理を繰り返す。

【0112】

図11はリモートコピーシステムに障害が発生した際に正記憶サブシステムが実行する処理の一例を示している。

【0113】

正記憶サブシステム104aは障害を検出すると（1101）、障害が発生している部位を特定して（1102）、管理コンソール117に対して障害を報告する（1103）。障害の報告を受けた管理コンソール117は、管理者からの指示に基づいて障害部位を閉塞させる（1104）。尚、障害部位の閉塞は管理コンソール117以外にも障害を検出した正記憶サブシステム104からも行う。

【0114】

障害部位を閉塞した後、正記憶サブシステムはリモートコピーの通信経路の状態を取得し、交替パスが存在するか否かをチェックする（1105）。交替パスが存在する場合にはリモートコピーの処理が継続可能であるため、正記憶サブシステム104は交替パスへのリモートコピー用パスの切り替えを行った後に障害処理を終える。

【0115】

交替パスが存在しない場合には正記憶サブシステムはリモートコピーペアの状態を変更する。尚ステップ904において、正記憶サブシステム104aが中間記憶サブシステム104cのキュー領域に空き領域がないと判断した場合であって、キュー領域に空き領域がない状態を障害の発生とみなす場合にも、交替パスが存在しない場合として扱う。

【0116】

交代パスが存在しない場合、正記憶サブシステム 1 0 4 が障害を検出していれば、自らペア状態を変更し(1106)、差分情報（ビットマップ）に副記憶サブシステム 1 0 4 bには格納されていないライトデータの位置情報を登録する（1107）。尚、別途管理コンソール 1 1 7 がペアの状態変更、差分情報（ビットマップ）の作成処理を正記憶サブシステムに指示しても良い。

【 0 1 1 7 】

ステップ1106によって、ペアの状態は障害検出前の正常状態（ペア状態）からペアが切り離されている状態（サスペンド）へ変更される。また、ビットマップは正記憶サブシステム 1 0 4 a上のメモリ 2 0 5 またはディスクドライブ 2 1 0 上に作成される。尚、ビットマップ作成開始時点の時刻、各グループの更新要求に関する要求番号等も記憶サブシステムに記憶される。

【 0 1 1 8 】

また、正記憶サブシステム 1 0 4 a は未反映のリモートコピー更新要求を先頭にビットマップを作成する。例えばリモートコピーシステムが図 6 の状態にあった時点でサスペンド状態となった場合には、正記憶サブシステム 1 0 4 a はビットマップを作成して D a t a 1 ～ 6 に関する差分情報をビットマップに格納し、これを初期状態とした後に、それ以後ホスト 1 0 2 から受信したライト要求についてビットマップを作成する（910）。

【 0 1 1 9 】

障害発生後も正記憶サブシステム 1 0 4 a に対して正ホスト 1 0 2 a からの業務が継続される場合は、ホスト 102a から正記憶サブシステム 1 0 4 a がライト要求とライトデータを受信する度に、ステップ1107の処理によって、正記憶サブシステム 1 0 4 a 上のビットマップへ差分情報が格納される。

【 0 1 2 0 】

尚、上記のステップ 1 1 0 1 からステップ 1 1 0 6 に示す処理は処理時間が十分短いためホスト 1 0 2 の I O 要求を正記憶サブシステムが受け付け、処理を継続したまま処理可能である。またステップ 1 1 0 7 に示す処理も全リモートコピーペア同時に実施すると時間を要するが、個々のリモートコピーペアに対しては十分短い時間で処理可能である。ステップ1107に示すビットマップの更新処理は

ホスト 102 からの I/O 要求に同期して実行することによりリモートコピーペア間では分散処理が可能となるので、ホスト 102 の I/O 処理を継続したまま処理できる。

【0121】

図 12 は副記憶サブシステム 104b が制御情報を取得し、これを更新する処理の一例を示す図である。

【0122】

まず副記憶サブシステム 104b は構成情報 106 に基づいて、正制御情報 631 を中間記憶サブシステム 104c の制御情報領域 113 から取得する (1201)。次にステップ 1001 と同様、制御情報取得が正常に行われたか、ハートビート情報が更新されているか等をチェックする (1202)。ステップ 1202 において障害を検出した場合には、副記憶サブシステムは図 13 に示す障害時処理を実施する (1209)。

【0123】

正常な状態の場合、即ち障害が検出されなかった場合には、副記憶サブシステムはまず更新要求の有無を調べる (1203)。副記憶サブシステムは、取得した正制御情報 631 の正先頭位置 619 と正サイズ 620 で示されるリモートコピー処理中の要求と、自装置が有する制御情報 107b の先頭位置 629 とサイズ 630 で示されるリモートコピー処理中の要求との差分から、前回自装置が制御情報 107b を更新した時点以降にキュー領域 114 に追加された更新要求を把握することができる。

【0124】

ステップ 1203 において新たな更新要求 (即ち、副記憶サブシステムが前回制御情報 107b を更新した後、新たにキュー領域に追加された更新要求) がない場合には、副記憶サブシステムは制御情報更新処理を終え一定期間待機する (1208)。

【0125】

更新要求がある場合には、副記憶サブシステム 104b は当該更新要求をキュー領域 114 から取得する (1204)。副記憶サブシステム 104b が取得可能なデータの容量が十分大きい場合には、副記憶サブシステム 104b は取得すべき更新要

求を全て一度にキュー 1 1 4 より取得する。

【 0 1 2 6 】

その後、副記憶サブシステム 1 0 4 b 内にて更新要求の内容を解析し、更新要求に従って正記憶サブシステムから送信されたデータを副記憶サブシステム 1 0 4 b 内のボリュームへ反映する（ 1 2 0 5 ）。即ち、更新要求に含まれるアドレス 7 0 7 が示す記憶領域に、更新要求に含まれるデータ 7 2 3 を格納する。

【 0 1 2 7 】

その後更新を行ったデータに対応する更新要求の分だけ、副記憶サブシステムが有する制御情報 1 0 7 b を更新し（具体的には先頭位置 629 と副サイズ 630 を更新し、）（ 1 2 0 6 ）、更新後の制御情報 107b の内容を用いて中間記憶サブシステム 1 0 4 c の制御情報領域 1 1 3 に格納されている副制御情報 632 を更新する（ 1 2 0 7 ）。

【 0 1 2 8 】

尚、ステップ 1 2 0 3 において更新要求のサイズが大きく、更新要求を一度に取得することができない場合、副記憶サブシステム 104b は更新要求を分割して取得し、処理を行う。まず副記憶サブシステム 1 0 4 b は更新要求取得のための領域を用意し更新要求を先頭位置から取得する。この領域は副記憶サブシステム 1 0 4 b が用意可能な領域、例えば 1 0 MB 等となる。

【 0 1 2 9 】

尚、この記憶領域に格納される最後の更新要求に関しては、ほとんどの場合取得できるのは最後の更新要求の途中までである。例えば 1 0 MB の領域が副記憶サブシステムに確保されており、ここに既に 9 MB 分の更新要求が取得済みとなっている場合を考える。副記憶サブシステムは次の更新要求を 9 MB の位置から自装置内の領域に読み込むわけだが、次の更新要求のサイズが 2 MB であった場合は、当該更新要求のうち最初の 1 MB 部分は取得できるが、残りの部分は取得できない。

【 0 1 3 0 】

このような場合副記憶サブシステムは、最初の 1 MB 部分のみ中間記憶サブシステムから更新要求を取得し、当該部分に対応するヘッダ情報を解析して先頭部

分に格納されているデータのみ副記憶サブシステム内のボリュームへ反映する。そして、処理した更新要求の先頭部分に関してのみ、制御情報 1 0 7 b の副先頭位置 6 2 9、副サイズ 6 3 0 を変更して、中間記憶サブシステム 1 0 4 c 上の制御情報 1 1 3 を更新する。尚、更新要求の残りの部分に関しては、副記憶サブシステムは次の更新要求取得時に処理する。

【 0 1 3 1 】

図 1 3 は障害発生時に副記憶サブシステム 1 0 4 b において実行される処理の一例を示す図である。障害時には副記憶サブシステム 1 0 4 b は可能な限り中間記憶サブシステム 1 0 4 c 上に残留する更新要求を取得する。正記憶サブシステム 1 0 4 a と中間記憶サブシステム 1 0 4 c は同期しているため、正記憶サブシステム 1 0 4 a のみに障害が起きた場合であって、中間記憶サブシステム 1 0 4 c にはアクセス可能な場合には、副記憶サブシステム 1 0 4 b は中間記憶サブシステム 1 0 4 c に格納されている更新要求を取得することにより、データ消失を防ぐことができる。

【 0 1 3 2 】

副記憶サブシステム 104b はまず、正記憶サブシステムに発生した障害を、例えば中間記憶サブシステム 1 0 4 c から取得した正制御情報 631 のハートビートから検出する (1 3 0 1)。そして、副記憶サブシステム 1 0 4 b は、中間記憶サブシステム 1 0 4 c の状態を調べる (1 3 0 2)。

【 0 1 3 3 】

中間記憶サブシステム 1 0 4 c に対してアクセスが出来ない場合にはこれ以上の更新要求の取得が不可能であるため、副記憶サブシステム 1 0 4 b はリモートコピーのペア状態をサスペンドへ変更し (1 3 0 8)、差分ビットマップを作成する (1 3 0 9)。

【 0 1 3 4 】

この場合ビットマップは障害発生時の時点から生成されることになる。図 6 の例においては、副記憶サブシステム 1 0 4 b は D a t a 2 までは更新要求を取得してリモートコピー処理を終了しているため、D a t a 2 以後の情報に関しビットマップへ差分情報を格納する。即ち副記憶サブシステム 1 0 4 b は、ビットマ

ップを作成後、D a t a 2 以降に副記憶サブシステム104bで更新されたの更新情報に関する差分情報をビットマップに格納する。

【0 1 3 5】

副記憶サブシステム 1 0 4 b にてデータが更新され、差分情報が作成される場合としては、例えばメンテナンス等のために正記憶サブシステムをサスペンド状態にし、副ホスト 1 0 2 b が副記憶サブシステムを用いて業務を実施する場合や、本実施例の様に正記憶サブシステムにおいて障害が発生した後に、副ホスト 1 0 2 b へ業務を移行し、副ホストが副記憶サブシステムに対して入出力要求を発行する場合がある。この際副記憶サブシステム 1 0 4 b の情報は D a t a 2 までは確定されており、副ホスト 1 0 2 b からのライトは D a t a 2 以降の更新情報として差分情報が保持される。

【0 1 3 6】

図 1 1 の説明で述べた様に、正記憶サブシステムにおいても副記憶サブシステムサスペンド時には、サスペンド後の更新情報について差分情報を保持している。このようにしてサスペンドの後はサスペンド時刻を起点に正記憶サブシステム 1 0 4 a、副記憶サブシステム 1 0 4 b の双方にて差分情報が保持されるので、図 1 4 で示すように障害回復後の際同期処理の際には、この差分情報を用いて再同期処理を実行することができる。

【0 1 3 7】

次にステップ 1 3 0 2 において、中間記憶サブシステム 1 0 4 c がアクセス可能であると判断された場合に関し説明する。まず副記憶サブシステム104bは、中間記憶サブシステム 1 0 4 c の制御情報領域 1 1 3 をチェックし、副記憶サブシステムが取得していない要求が存在するか調べる（1 3 0 3）。中間記憶サブシステムに要求が存在する場合には、副記憶サブシステム104bはキュー領域114から更新要求を取得し、取得した更新要求に従って、副記憶サブシステム 1 0 4 b のボリュームにデータを反映する（1 3 0 4）。そして副記憶サブシステム 1 0 3 b は、反映した更新要求だけ制御情報 1 0 7 b の先頭位置をインクリメントし（1 3 0 5）、再び中間記憶サブシステム104cの制御情報領域113を参照することにより、次の要求の有無をチェックする（1 3 0 6）。

【0138】

次の更新要求の有無は、（１）副記憶サブシステムが更新要求のヘッダ部分のみを取得して解析した後、更新要求の本体であるデータ部分及び末尾情報を取得する方法と、（２）更新要求から一定サイズ分のデータを副記憶サブシステム 104b が取得し、副記憶サブシステム内で解析する方法とが考えられる。本実施例では（１）を用いて説明するが、本発明ではどちらでも実現可能である。

【0139】

副記憶サブシステムはまず、キュー要求のヘッダ部分を取得して、リモートコピーのグループ ID、要求 ID、及びタイムスタンプを調べる。副記憶サブシステムはグループ ID、キュー ID を用いて論理的に矛盾がないかを解析し、要求 ID を用いて直前の更新要求との連番であるかをチェックし、タイムスタンプを調べて当該タイムスタンプが直前の更新要求のタイムスタンプよりも大きい値かをチェックする。副記憶サブシステムは、調査の結果論理的不整合がなければ更新要求が存在すると判断し、論理的不整合が発生する場合には更新要求が存在しないものと判断する。

【0140】

更新要求が存在する場合には、副記憶サブシステムは可能であれば更新要求全体を中間記憶サブシステム 104c から取得し、各ブロックの検査と末尾情報の検査をした後、検査結果が正常な場合に副記憶サブシステム 104b のボリュームにデータを反映する（1307）。正記憶サブシステム 104a が中間記憶サブシステム 104c へ更新要求を送信している最中に障害が発生した場合には、最後に書き込まれた更新要求に関しては書き込みの途中となってしまう可能性がある。このため、各ブロックの検査や末尾情報の検査により、更新要求の一貫性を確保する。

【0141】

その後、副記憶サブシステムは、再度アドレスをインクリメントし（1305）、中間記憶サブシステム 104c に次の更新要求が存在するかどうかを再度チェックする（1306）。副記憶サブシステムは中間記憶サブシステムに更新要求が存在しなくなるまで、ステップ 1305 からステップ 1307 までの処理を

繰り返す。

【0142】

以上の処理を図6を用いて説明する。まず副記憶サブシステム104bは制御情報113を参照て、キュー領域114からData3, Data4をステップ1304にて中間記憶サブシステムからまとめて取得する。その後ステップ1305から1307においてData5及びData6各々に関し、更新要求ごとに各々ヘッダ、末尾情報等の解析を行い、論理的不整合が発見されなければData5及びData6を取得する。そして、副記憶サブシステムは、Data6の次の更新要求を取得しようとする際に、ヘッダ、データまたは末尾情報のいずれかで情報の不整合を発見し、更新要求の取得を終了する。

【0143】

図14は正記憶サブシステム104aと副記憶サブシステム104b間の再同期処理の一例を示す図である。

【0144】

まず管理者は管理コンソール117にて各装置の状態や、装置間の接続状態を取得し再同期が可能な状態かを調べる(1401)。データ処理システム内の経路の一部や記憶サブシステム104が使用不可能となっている等の理由で、再同期が不可能な状況の場合は(1402)、エラー処理が実行される(1414)。

【0145】

管理者はまた、管理コンソールを用いて、データ処理システムの論理的な整合性もチェックする。管理者は、管理コンソールを用いて、全記憶サブシステム104について、リモートコピーのグループに属するボリュームが使用可能かを確認し、各グループや各ボリュームに関して障害の検知時刻やビットマップの開始時刻を比較する。管理者は再同期処理に用いられるビットマップの開始時刻を比較することで、データ処理システムの論理的な整合性を確認できる(1403)。

。

【0146】

副記憶サブシステム104bは、例えば図13のステップ1302において中間記憶サブシステムにアクセスできないと判断された場合(1310)の様に、状況により

正記憶サブシステム 1 0 4 a と同等の情報が格納されていない場合がある。従って上述の様に、正記憶サブシステム 1 0 4 a では副記憶サブシステム 1 0 4 b によって取得され処理されたことが確認されていない更新要求に関してはビットマップを作成する必要がある。

【 0 1 4 7 】

例えば、図 6 の例においては、正記憶サブシステム 1 0 4 a は D a t a 1 からビットマップの取得を開始するべきである。何らかの論理的不整合において D a t a 6 以降の時刻からしかビットマップを取得していない場合には、正記憶サブシステムと副記憶サブシステムとの間で再同期処理をした際にデータの不整合が発生する。この場合 D a t a 3、D a t a 4、D a t a 5、及び D a t a 6 が再同期できないという不整合が発生する。つまり、正記憶サブシステム 1 0 4 a のビットマップ作成時刻（差分情報取得開始時刻）が副記憶サブシステム 1 0 4 b のビットマップ作成時刻（差分情報取得開始時刻）より新しい場合には（1 4 0 3）、データの不整合が発生するので、エラー処理が実行される（1 4 1 4）。

【 0 1 4 8 】

データの不整合が生じていないと判断された場合には、再同期処理としてまず、管理コンソール 1 1 7 上で取得したシステムの状態を元に、管理者が各リモートコピーのペアを再同期すべく経路等を生成する（1 4 0 4）。この際経路は必ずしも中間記憶サブシステム 1 0 4 c を介する必要はなく、正記憶サブシステム 1 0 4 a と副記憶サブシステム 1 0 4 b を直接接続してもよい。

【 0 1 4 9 】

次に管理者は管理コンソール内に新しい構成情報 1 0 6 を作成し、管理コンソール 1 1 7 は新構成情報 1 0 6 を各記憶サブシステム 1 0 4 へ送付する（1 4 0 5）。各記憶サブシステム 1 0 4 は送付された新構成情報 1 0 6 を元に指示された経路を介して指示された装置への接続を試みる。

【 0 1 5 0 】

次に再同期の要否が判定される。正記憶サブシステム 1 0 4 a または副記憶サブシステム 1 0 4 b のいずれかに障害が発生しデータが消失した場合には再同期はせずに、初期コピーを行う（1 4 1 3）。

【0151】

再同期が必要な場合にはどちらの記憶サブシステム 104 へデータ内容を合わせこむかを判定する (1407)。メンテナンス等で一時的に副記憶サブシステム 104 b に業務が移行した場合や、障害の発生により正記憶サブシステム 104 a から副記憶サブシステム 104 b へ業務を移行した場合には、副記憶サブシステム 104 b から正記憶サブシステム 104 a へ再同期が必要になる。また障害の発生により、副記憶サブシステムでは更新データを反映させることなく、正記憶サブシステム 104 a において業務を継続した場合には、正記憶サブシステム 104 a から副記憶サブシステム 104 b へのデータの再同期が必要になる。即ち、通常は業務を移管したシステム側、若しくは業務を継続して実行していたシステム側にデータ内容を合わせこむ。

【0152】

正記憶サブシステム 104 a から副記憶サブシステム 104 b へ再同期を行う場合には、副記憶サブシステムが有するビットマップを副記憶サブシステム 104 b から正記憶サブシステム 104 a へ送付し (1408)、正記憶サブシステム 104 a において正記憶サブシステム 104 a のビットマップと副記憶サブシステム 104 b のビットマップを統合し再同期用のビットマップを生成する。ビットマップの統合は、具体的には両方のビットマップ上に示されている各ビットの論理和 (OR) を計算することで得られる。

【0153】

そして、再同期用のビットマップに従って、副記憶サブシステム内に未だ反映されていない更新要求が正記憶サブシステムから副記憶サブシステムに送信され、最同期処理が実行される (1409)。尚、再同期処理開始後は、再同期中においても“Copy on write”等の技術を用いることで、記憶サブシステム 104 はホスト 102 から受信する I/O 処理要求の処理を再開することができる (1410)。

【0154】

尚、副記憶サブシステム 104 b から正記憶サブシステム 104 a への再同期処理は、ビットマップの送付方向 (1411)、更新要求の送信方向 (1412

）が逆となるだけで、その他は正記憶サブシステムから副記憶サブシステムへの再同期処理と同様の処理によって実行される。

【0155】

－第2の実施形態－

図15にデータ処理システムの別の実施形態を示す。図15では中間記憶サブシステム104c上にはリモートコピーキュー114が存在し、構成情報106及び制御情報107は中間記憶サブシステム104cには格納されない。構成情報106は管理コンソール117から正記憶サブシステム及び副記憶サブシステムに送付される。制御情報107は中間記憶サブシステム104cを介さずに正記憶サブシステム104aと副記憶サブシステム104bとの間で直接やり取りされる。

【0156】

第1の実施形態と同様に、正ホスト102aから正記憶サブシステム104aへのデータのライトと同期して、正記憶サブシステム104aから中間記憶サブシステム104cへ更新要求が送信される。一方制御情報107は、ホスト102から発行されるライト要求とは非同期に正記憶サブシステム104aから副記憶サブシステム104bへ直接送信される。副記憶サブシステム104bは受信した制御情報107bを参照して中間記憶サブシステム104cから更新要求を取得し、更新要求から取得したライトデータを副記憶サブシステムのディスクに格納することによって、リモートコピーを実現する。

【0157】

正記憶サブシステム104aと副記憶サブシステム104b間で制御情報を直接送受信することで、中間記憶サブシステム104cを介するために必要となるデータ一貫性チェックや送信遅延を削減することができる。

【0158】

－第3の実施形態－

図16にはデータ処理システムの他の一例として、中間記憶サブシステム104cを複数有するシステムの例を示す。本実施形態では中間サイト101が中間サイト101A、中間サイト101Bと複数存在する。図16に示すデータ処理

システムにおいては、（１）リモートコピーグループ毎に使用する中間記憶サブシステム 1 0 4 c を割り当てる、若しくは（２）任意のリモートコピーグループのリモートコピー処理において複数の中間記憶サブシステム 1 0 4 c を使用する、２種類の実施形態が存在する。

【 0 1 5 9 】

（１）の実施形態の場合、各リモートコピーグループに着目すると実施の形態 1 と同様の処理となる。

【 0 1 6 0 】

（２）の実施形態の場合、正記憶サブシステム 1 0 4 a が更新要求を複数の中間記憶サブシステム 1 0 4 c に振り分けて送信する処理を実施する。この際各更新要求の ID はリモートコピーグループ内、及び中間記憶サブシステム 1 0 4 c 間に渡り唯一の連続番号である。副記憶サブシステム 1 0 4 b は複数の中間記憶サブシステム 1 0 4 c から取得した更新要求を整理した後に、更新要求に従ってデータを自装置が有するディスク装置のボリューム内に反映する。

【 0 1 6 1 】

本実施形態においては障害時の処理においても特徴がある。本実施形態のように複数の中間記憶サブシステム 1 0 4 c を有することで、任意の一つの記憶サブシステム 1 0 4 が故障した場合でも、データ処理システムの構成情報を再構築すれば、交替パス制御が可能である。交替パス制御処理を行う際には障害が発生した中間記憶サブシステム 1 0 4 c 上に滞留している更新要求を、正記憶サブシステムから副記憶サブシステムに、別の正常な中間記憶サブシステム 1 0 4 c を経由して送信しなおす必要がある。再送される更新要求は正記憶サブシステムが過去に障害が発生した中間記憶サブシステムに書き込んだ更新要求であるから、再送される更新要求の ID は、正記憶サブシステム 104a が直前に中間記憶サブシステムに送信した更新要求の ID とは連続しない可能性がある。しかし、副記憶サブシステム 1 0 4 b 側で中間記憶サブシステムから取得した更新要求を要求 ID 順に整理した後、ボリュームへのデータ格納処理を実行すれば問題は生じない。

【 0 1 6 2 】

－ 第 4 の実施形態 －

図 17 は交替パスを有するデータ処理システムの一例を示す。図 17 に示すシステムには、正記憶サブシステム 104 a と副記憶サブシステム 104 b を結合するリモートコピーリンク 1701 が存在する。リモートコピーリンク 1701 は (1) 中間記憶サブシステム 104 c の障害時にリモートコピーの交替パスとして、若しくは (2) 制御情報通信用パスとして使用可能である。

【0163】

(1) 中間記憶サブシステム 104 c に障害が発生して中間記憶サブシステム経由ではリモートコピーが実行できなくなった場合には、正記憶サブシステムと副記憶サブシステムは、代わりにリモートコピーリンク 1701 を用いてリモートコピーを実行することができる。リモートコピーリンク 1701 を障害時の交替パスとして使用する場合は、正記憶サブシステム 104 a と副記憶サブシステム 104 b 間は長距離であるため、非同期リモートコピーが用いられる。

【0164】

(2) 正常時においてはリモートコピーリンク 1701 を用いて制御情報 107 を非同期通信にて送受信することもできる。即ち、正記憶サブシステム 104 a から副記憶サブシステム 104 b へ、リモートコピーリンク 1701 を使用し制御情報 107 を送付することができる。また、更新要求に関するデータ以外のメッセージも、正記憶サブシステムと副記憶サブシステムの間でリモートコピーリンク 1701 を使用して非同期で送受信できる。正記憶サブシステム 104 a と副記憶サブシステム 104 b 間で、中間記憶サブシステムを介することなく直接通信することで、中間記憶サブシステム 104 c を介するために必要となる、データ一貫性チェックや送信遅延を削減することが可能である。

【0165】

－第 5 の実施形態－

図 18 には副記憶サブシステム 104 b を複数有するデータ処理システムの一例を示す。副記憶サブシステム 104 b が複数個存在する場合には構成情報 106 や制御情報 107 も、複数の副記憶サブシステム 104 b 各々について必要となる。

【0166】

具体的には正記憶サブシステム 104 a から中間記憶サブシステムを介して副記憶サブシステム 104 b へ送信される更新要求に関しては一対多の制御が必要になり、中間記憶サブシステムに格納される副制御情報 632 及び副情報 612 は副記憶サブシステム 104 b の個数分必要になる。

【0167】

副記憶サブシステム 104 b は第 1 の実施形態と同様の制御を行う。正記憶サブシステム 104 a 側の図 10 に示す処理が変更となる。正記憶サブシステム 104 a は複数の副記憶サブシステム 104 b の副制御情報 632 を取得し、処理 1003 にて全副記憶サブシステム 104 b の副制御情報 632 を比較する。比較の結果、最も更新が遅い（古い）副記憶サブシステム 104 b の先頭位置を新先頭位置として新しい制御情報 107 a を生成する。また障害時においても正記憶サブシステムは同様に処理し、処理 1107 のビットマップ生成において差分情報取得開始時刻を更新がもっとも遅い副記憶サブシステム 104 b に合わせる。これにより正記憶サブシステム 104 a と副記憶サブシステム 104 b 間で有効な差分情報を保持できるようになる。

【0168】

副記憶サブシステム 104 b から正記憶サブシステム 104 a へ送信する更新要求に関しては、一対一対応になるため、第 1 の実施形態と同様の制御が実行される。

【0169】

以上に説明したように、n サイト間で実行されるリモートコピー処理におけるコスト、処理の複雑さを削減するために、正記憶サブシステムと副記憶サブシステムとを接続する中間サイトには I/O 処理を実施する通常の間記憶サブシステムを配置し、リモートコピーに係わる処理は正サイトと正記憶サブシステムと副サイトの副記憶サブシステムが行う。即ち、中間記憶サブシステムはリモートコピー処理時には正記憶サブシステム若しくは副記憶サブシステムからの I/O 要求に応じてリード若しくはライトの処理を実行する通常の記憶サブシステムである。

【0170】

上記構成において正記憶サブシステムと副記憶サブシステムは中間記憶サブシステムを介してリモートコピーの処理を実行する。正記憶サブシステムと副記憶サブシステムは中間記憶サブシステム内のリモートコピーキュー 114 に、データやメッセージをリードしたりライトしたりすることによって、当該データやメッセージを正記憶サブシステムと副記憶サブシステム間でやり取りする。

【0171】

正記憶サブシステムがホストからライト要求と、ライトデータを受信すると、ライトデータ及びライトデータの格納位置等ライト要求に含まれる制御情報を有する更新要求を中間記憶サブシステムに送信し、その後にホストへ応答を返送する(同期リモートコピー)。またリモートコピーキュー 114 の進捗を示す先頭位置等のポインタや更新要求のサイズ等が含まれるキュー制御情報も正記憶サブシステムは中間記憶サブシステムに送信する。制御情報の更新はホストのライト要求とは同期でも非同期でもかまわないが性能向上のため一定間隔毎に非同期に更新するほうが好ましい。尚、中間記憶サブシステムを介することなく、正記憶サブシステムと副記憶サブシステム間を直接接続して制御情報を直接やり取りしてもかまわない。

【0172】

副サイトの副記憶サブシステムは中間記憶サブシステムに書き込まれた正制御情報を読み取り、正制御情報に基づいて中間記憶サブシステム上のキューからライトデータを含む更新要求を取得する。尚副記憶サブシステムは正記憶サブシステムから中間記憶サブシステムへの更新要求の送信とは非同期に、更新要求の取得を行う(非同期リモートコピー)。そして取得した更新要求に含まれる位置情報に基づいて、取得した更新要求に含まれるライトデータを副記憶サブシステム内のディスクに格納する。

【0173】

上記の処理において更新要求や制御情報のやりとりは、正記憶サブシステム及び副記憶サブシステムがリード若しくはライト等のコマンド中間記憶サブシステムに発行し、中間記憶サブシステムがこれに応じてリード若しくはライト処理を実行することによって実現される。このため、中間記憶サブシステムはリモート

コピーに係わる機能、プログラム等を必要とせず、低コストでnサイトリモートコピーを実現することができる。中間サイトの記憶サブシステムとして、より小型の記憶サブシステムサブシステムや、JBOD (Just a Bunch Of Devices)を使用すると一層の低価格化が可能である。

【0174】

また、中間記憶サブシステムには更新要求や制御情報を格納するための記憶領域は必要となるが、中間記憶サブシステムが正記憶装置システムや副記憶装置システムが有するデータのコピーを保持し続ける必要はないため、中間記憶サブシステムに必要な記憶容量は正記憶サブシステムや副記憶サブシステムが有する記憶容量より少なく済む。従って、より少ない記憶容量で、nサイト間でのリモートコピー処理を実行することができる。

【0175】

また中間記憶サブシステムは一般に、正記憶サブシステム及び副記憶サブシステムの双方との間でリモートコピー処理を実行する必要があるため、処理負荷が高くなってしまいが、本発明の実施形態においては、中間記憶サブシステムで実行される処理が単純化され、中間記憶サブシステムは正記憶サブシステム若しくは副記憶サブシステムから発行されるI/O要求を処理すれば良いので、中間記憶サブシステムの負荷を低減することができる。更に、中間記憶サブシステムについてはリモートコピーの進捗やステータスのチェック等の処理をする必要はないため、リモートコピーに要する監視・管理処理が軽減される。

【0176】

【発明の効果】

本発明によれば、低コストでnサイト間でのリモートコピーを実行することができる。また、本発明によれば、nサイト間でリモートコピーを実行するための処理に要する負荷容量を低減することができる。

【図面の簡単な説明】

【図1】 本発明が適用されるデータ処理システムの一例を示す図である。

【図2】 記憶サブシステムの構成例を示す図である。

【図3】 中間記憶サブシステムが有する記憶領域の一例を示す図である。

【図 4】 中間記憶サブシステムに格納されるデータの構造の一例を示す図である。

【図 5】 保護データの一例を示す図である。

【図 6】 中間記憶サブシステムが有する情報の構成例を示す図である。

【図 7】 要求のデータ構造の一例を示す図である。

【図 8】 初期設定の手順の一例を示す図である。

【図 9】 正記憶サブシステムにおける更新処理の一例を示す図である。

【図 10】 制御情報更新処理の一例を示す図である。

【図 11】 障害発生時に正記憶サブシステムにおいて実行される処理の一例を示す図である。

【図 12】 副記憶サブシステムにおける更新処理の一例を示す図である。

【図 13】 障害発生時に副記憶サブシステムにおいて実行される処理の一例を示す図である。

【図 14】 再同期処理の一例を示す図である。

【図 15】 本発明が適用されるデータ処理システムの他の一例を示す図である。

【図 16】 本発明が適用されるデータ処理システムの他の一例を示す図である。

【図 17】 本発明が適用されるデータ処理システムの他の一例を示す図である。

【図 18】 本発明が適用されるデータ処理システムの他の一例を示す図である。

【符号の説明】

102…ホスト

104…記憶サブシステム

106…構成情報

107…制御情報

108…更新情報

112…構成情報領域

113…制御情報領域

114…キュー領域

117…管理コンソール

201…コントローラ

2 0 2 …CPU

2 0 3 …ホストアダプタ

2 0 4 …ネットワークコントローラ

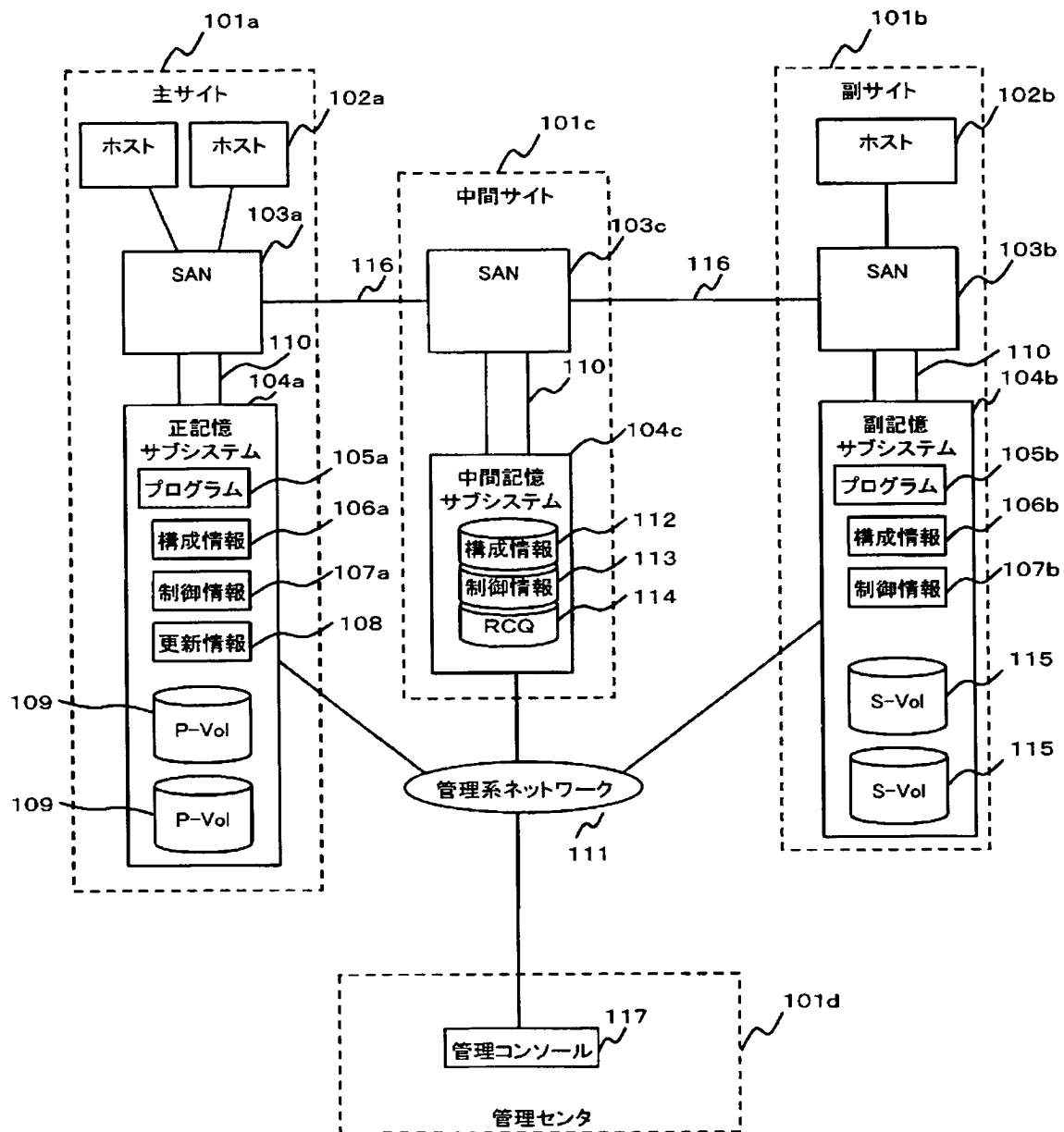
2 0 5 …メモリ

2 1 0 …ディスク装置

【書類名】 図面

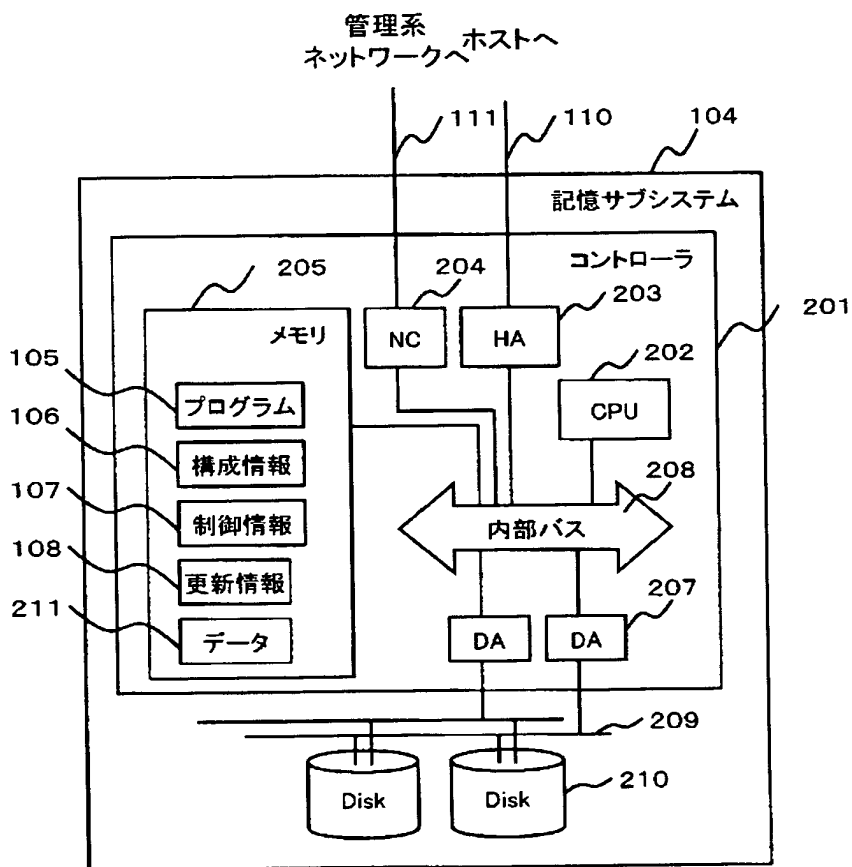
【図 1】

図 1



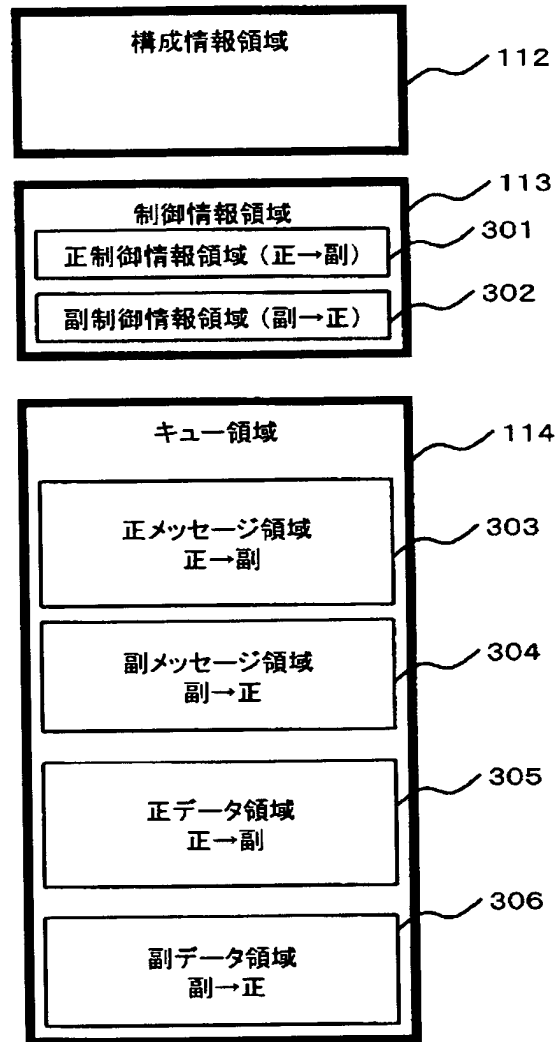
【図 2】

図 2



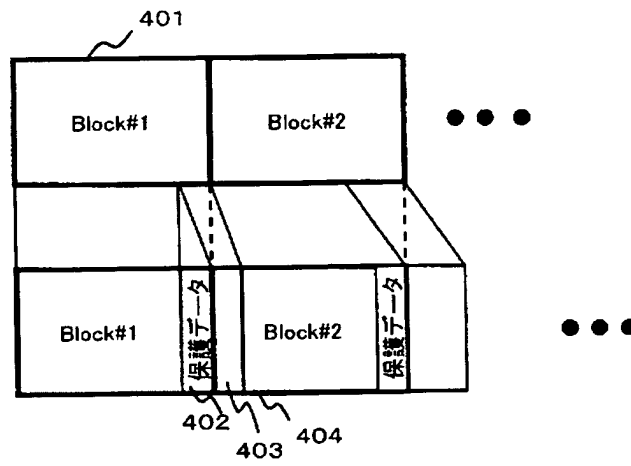
【図 3】

図 3

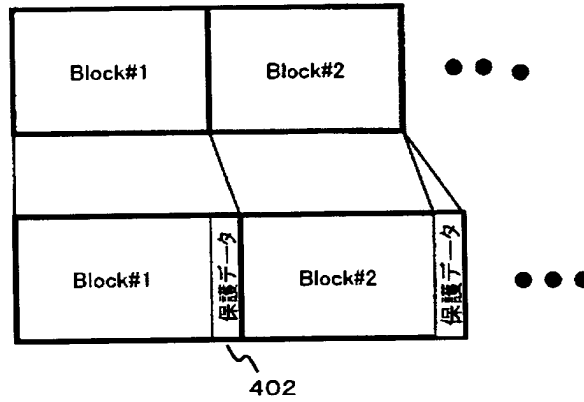


【図 4】

図 4



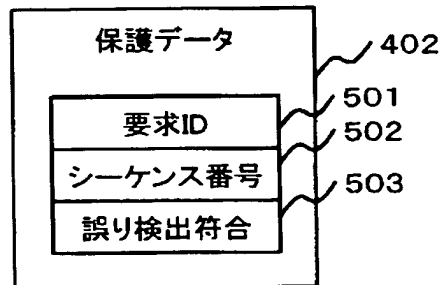
(a) 中間記憶サブシステムにてブロックサイズを変更なし



(b) 中間記憶サブシステムにてブロックサイズを変更

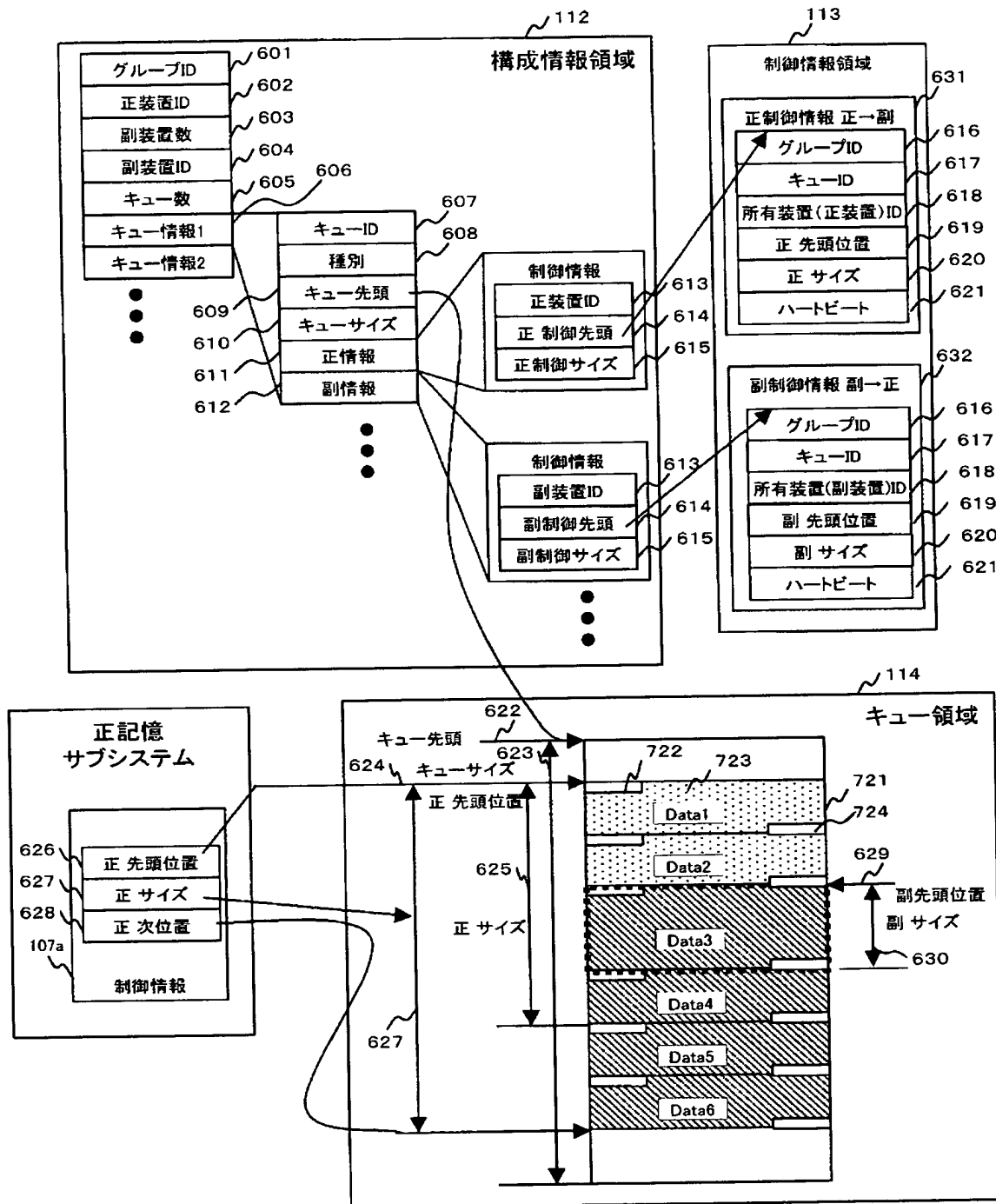
【図 5】

図 5



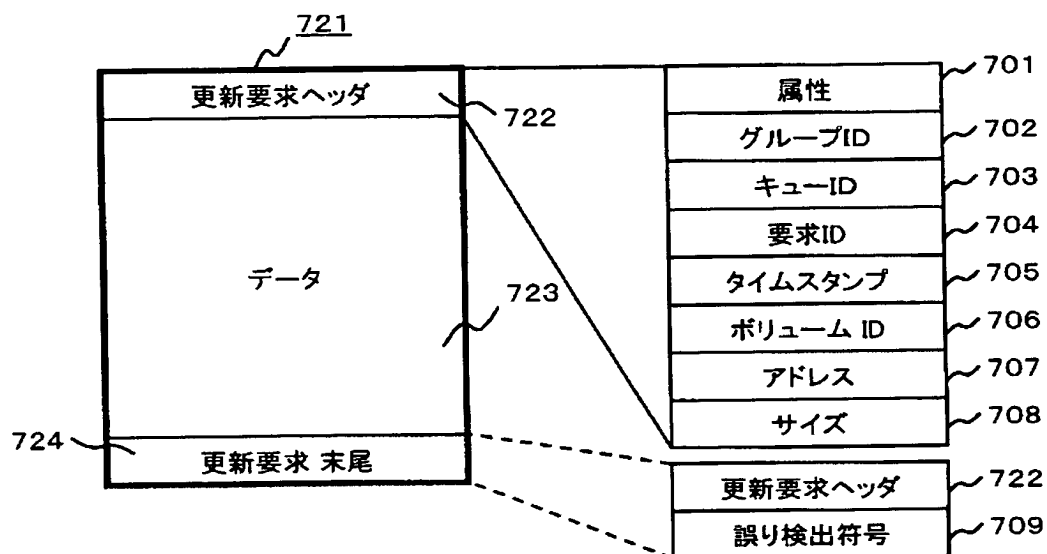
【図 6】

図 6



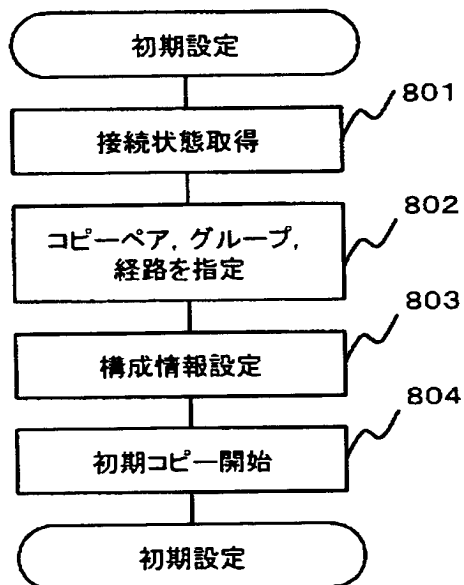
【図 7】

図 7



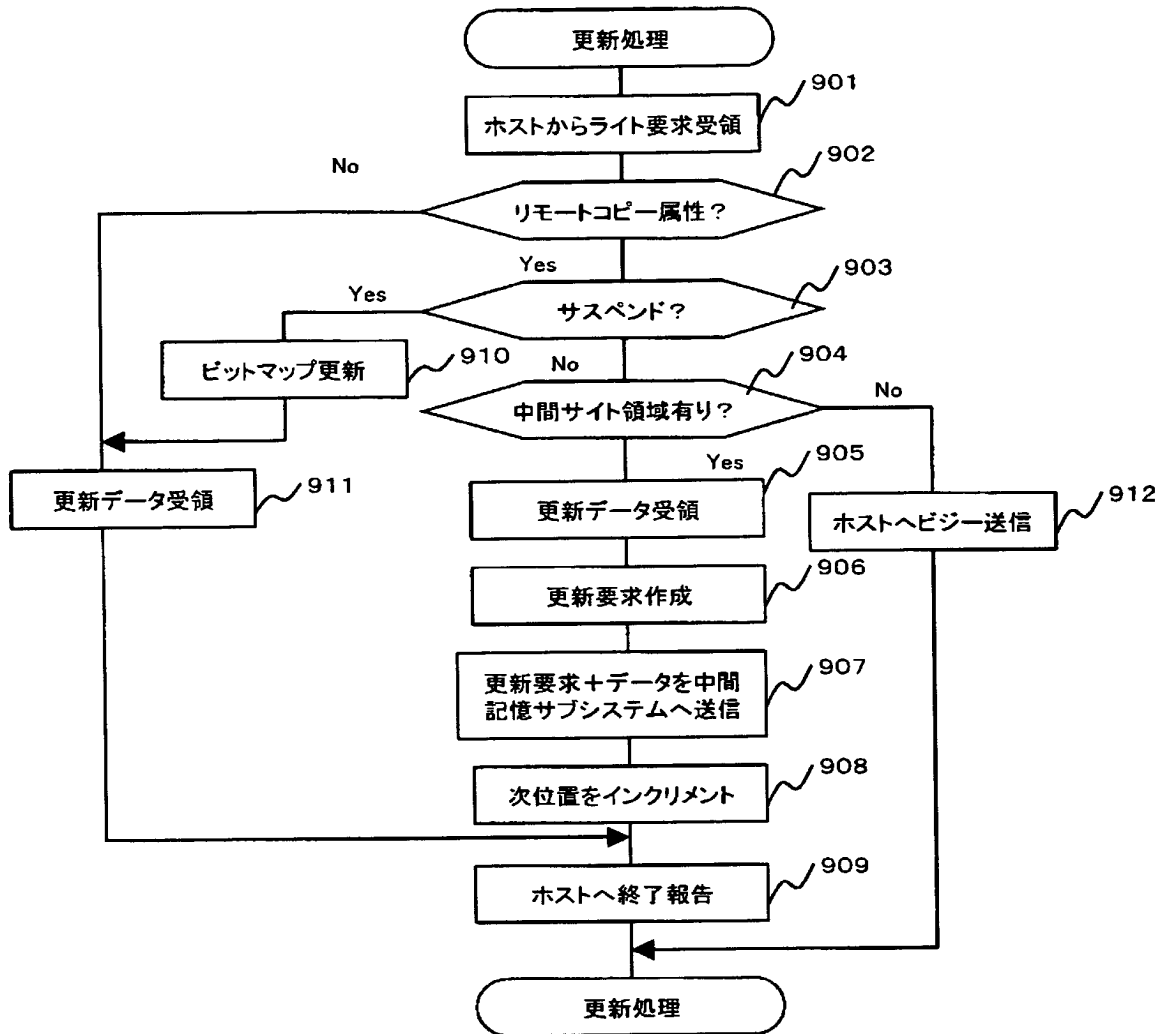
【図 8】

図 8

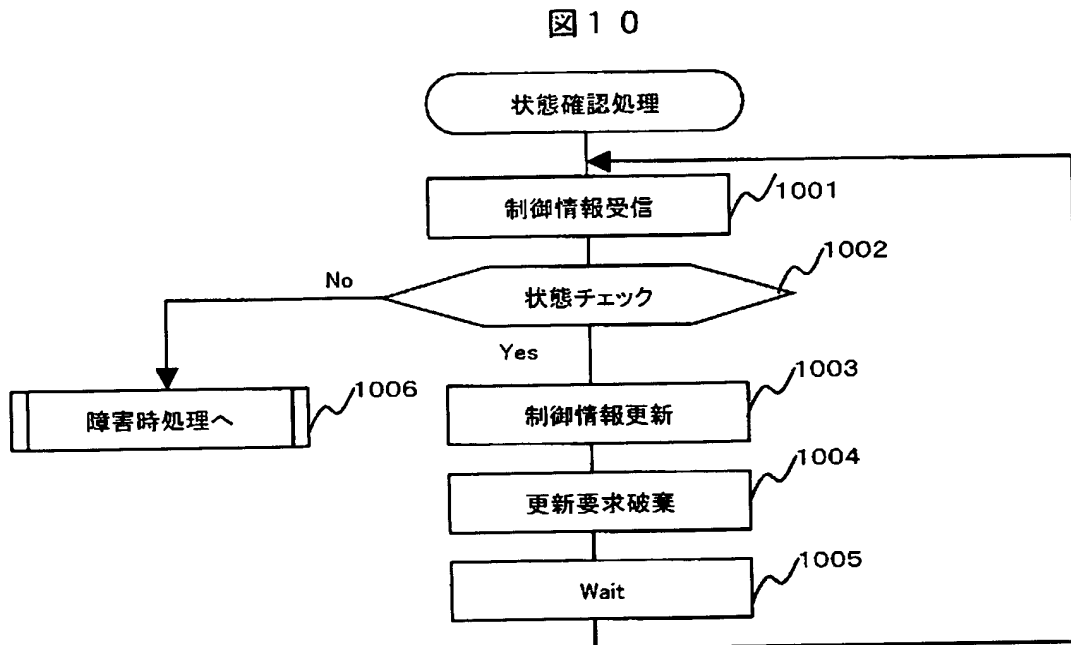


【図 9】

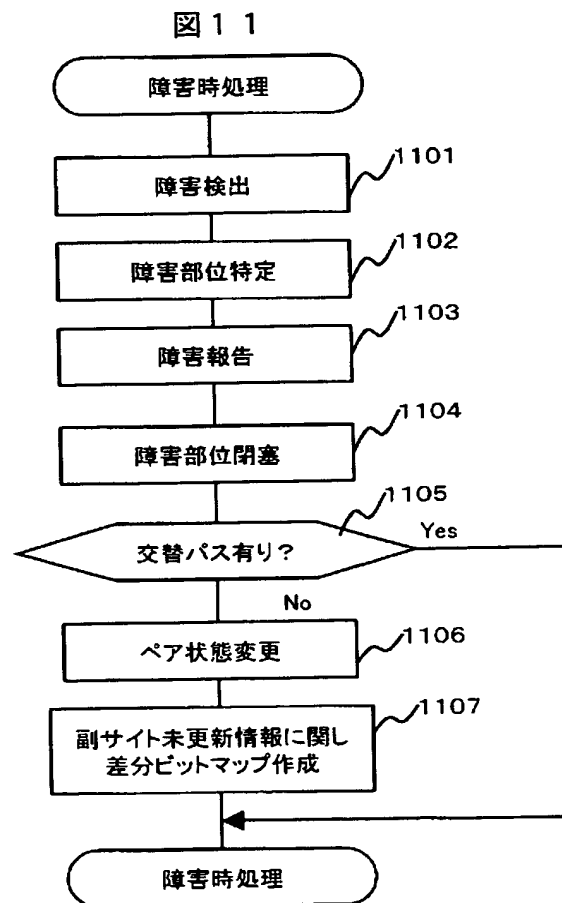
図 9



【図 10】

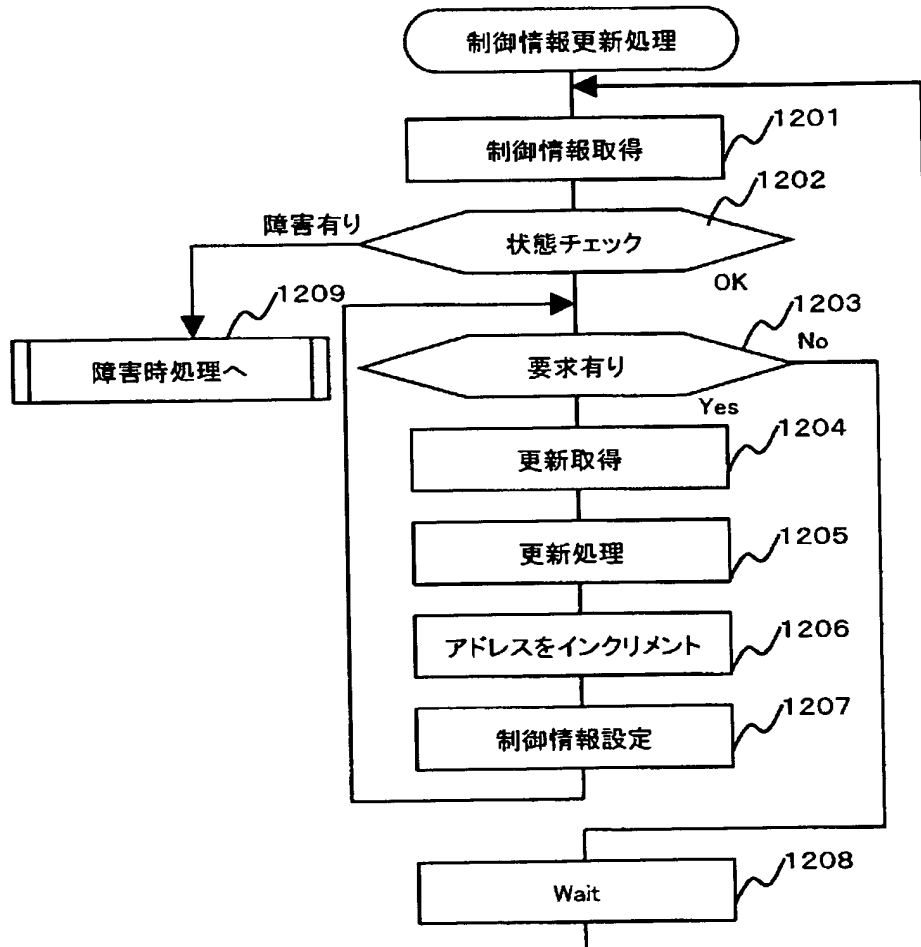


【図 11】



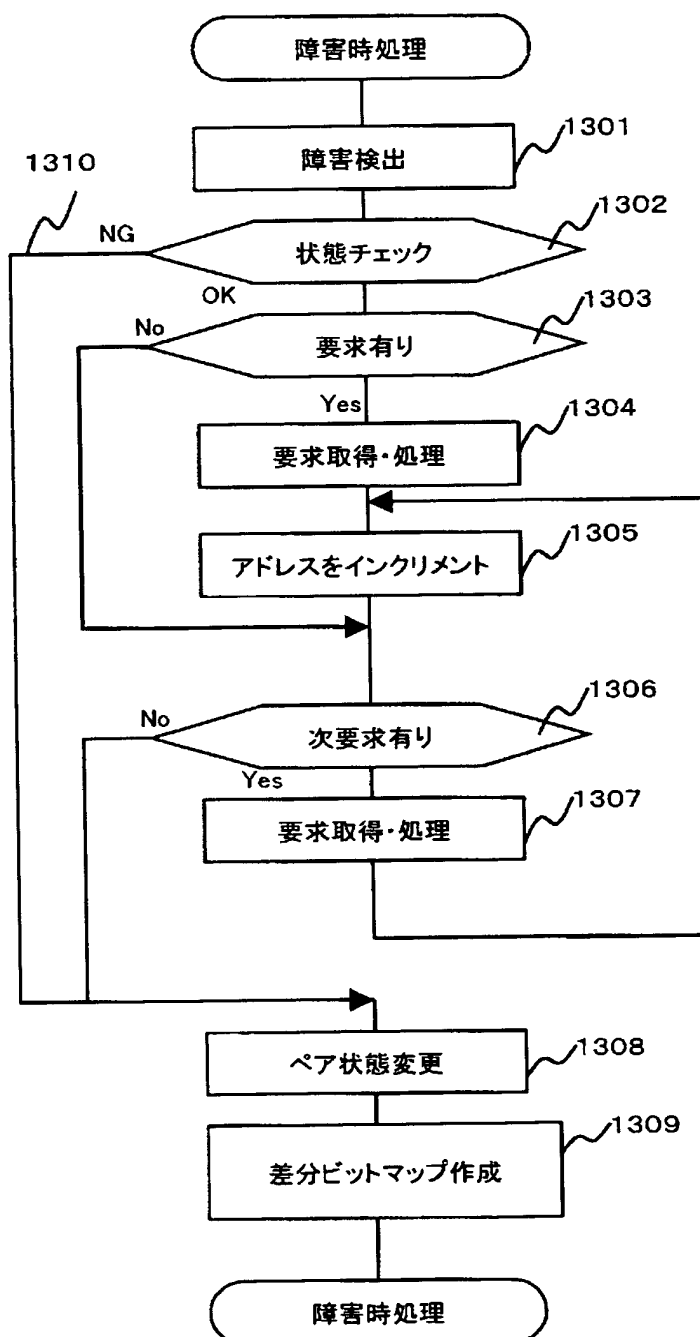
【図 12】

図 12

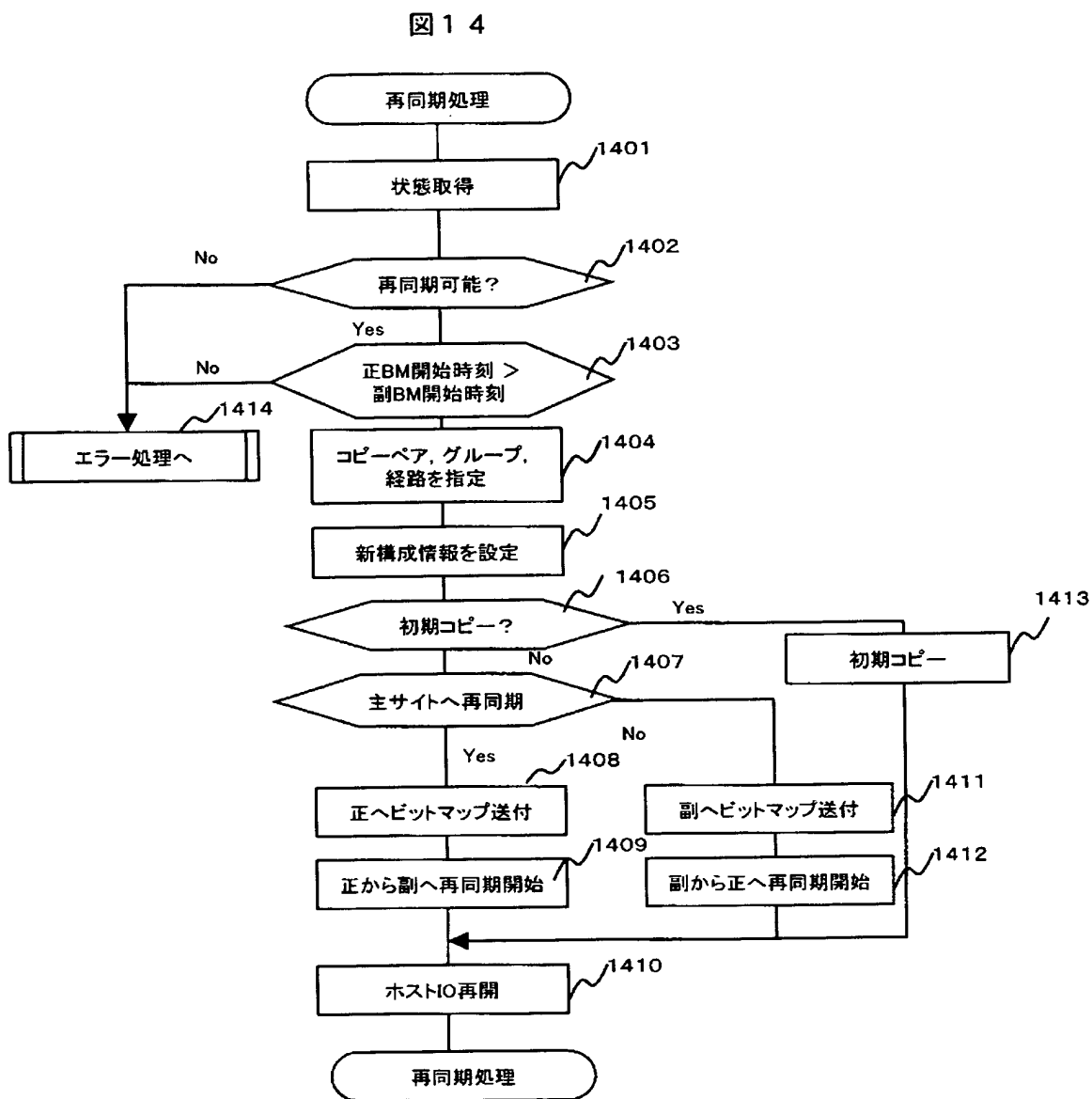


【図13】

図13

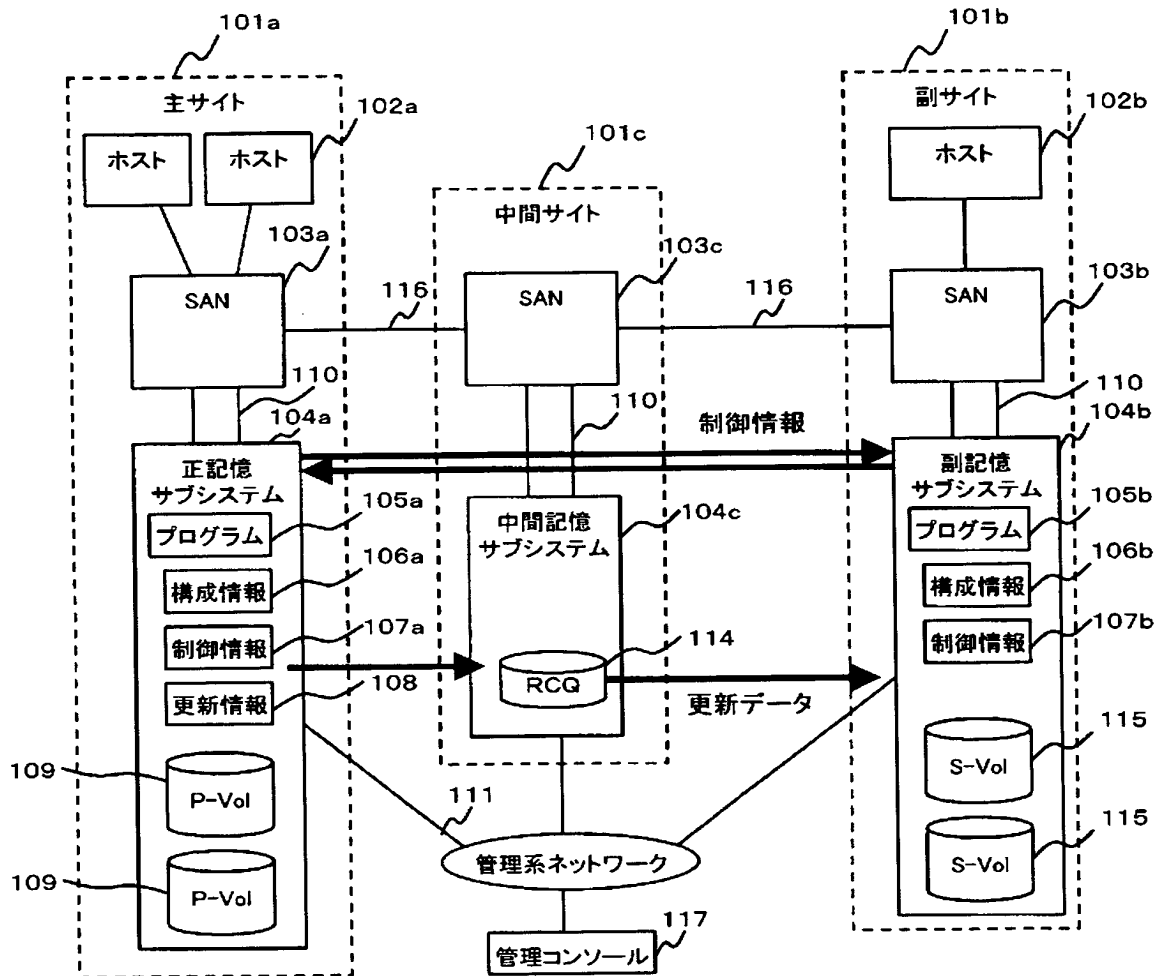


【図 14】



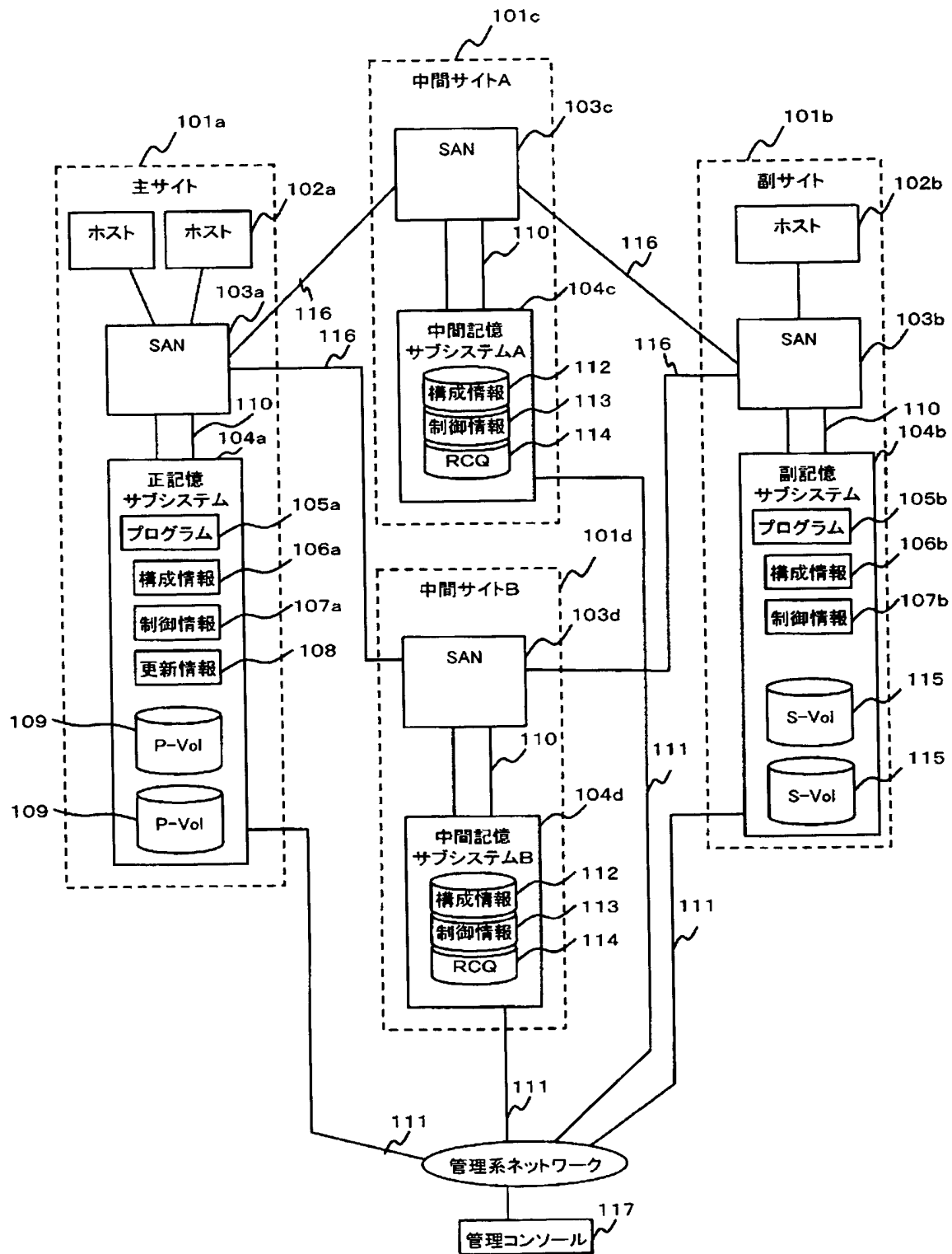
【図 15】

図 15



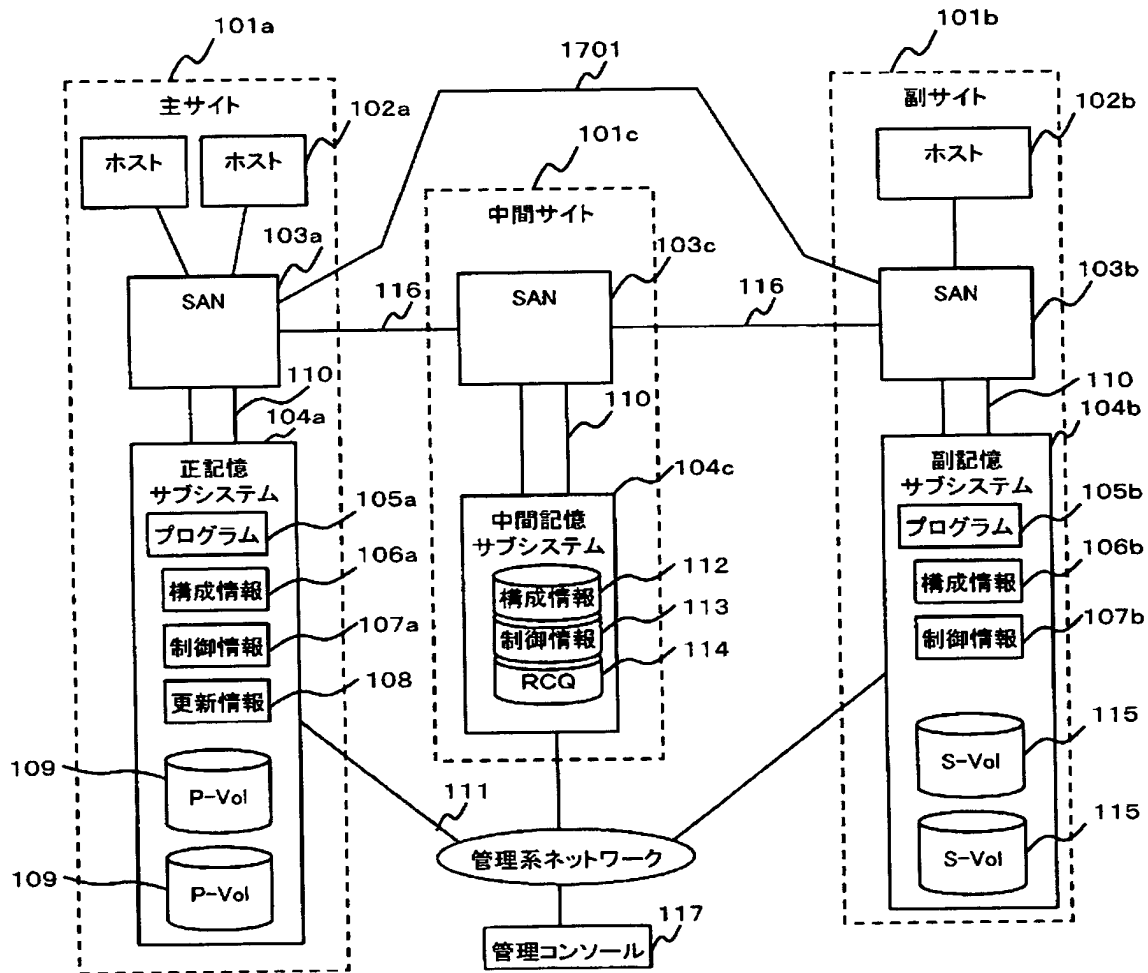
【図 16】

図 16



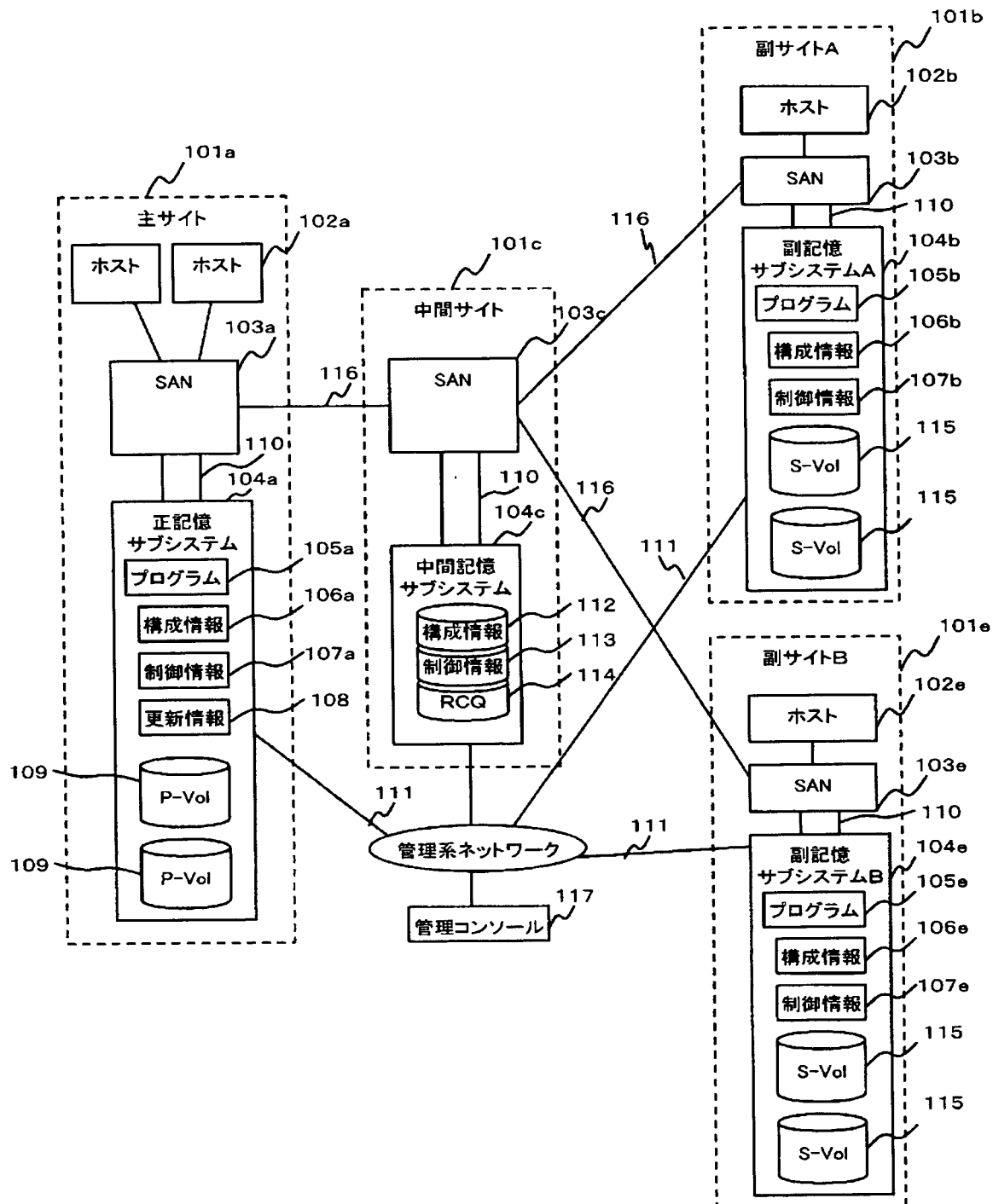
【图 17】

图 17



【図18】

図18



【書類名】 要約書

【要約】

【課題】

低コストで、処理負荷が小さいnサイトリモートコピーを実現する。

【解決手段】

第1の記憶装置システムと第2の記憶装置システムを第3の記憶装置システムを経由して接続する。リモートコピー処理を実行する際、第1の記憶装置システムは計算機から受信したライト要求に応じ、ライトデータとライトデータが書き込まれる格納位置を示すアドレス情報とを有するジャーナルを第3の記憶装置システムに送信し、第3の記憶装置に書き込みを要求する。第2の記憶装置システムは、第1の記憶装置システムによって発行されるジャーナルの格納位置を含む制御情報を受信し、制御情報に基づいて第3の記憶装置からジャーナルをリードする。そして第2の記憶装置システムは、ジャーナルに含まれるアドレス情報に従って、ジャーナルに含まれるライトデータを、第2の記憶装置システム内のディスクに書き込む。

【選択図】 図1

認定・付加情報

特許出願の番号	特願 2 0 0 3 - 2 0 5 6 1 7
受付番号	5 0 3 0 1 2 8 5 1 3 2
書類名	特許願
担当官	第七担当上席 0 0 9 6
作成日	平成 1 5 年 8 月 5 日

< 認定情報・付加情報 >

【提出日】 平成15年 8月 4日

特願 2 0 0 3 - 2 0 5 6 1 7

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 5 1 0 8]

1 . 変更年月日

1 9 9 0 年 8 月 3 1 日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台 4 丁目 6 番地

氏 名

株式会社日立製作所